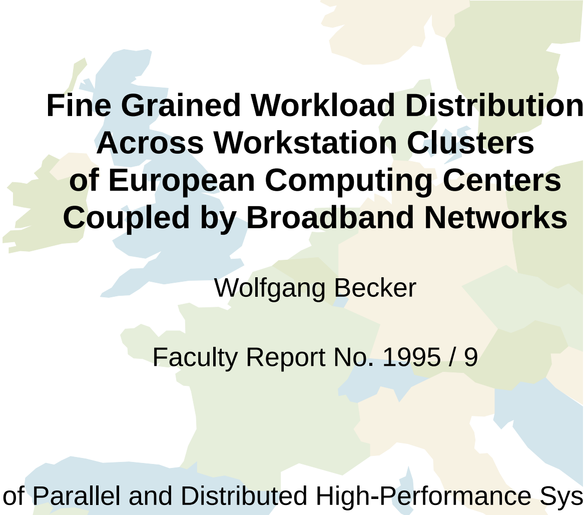
Abstract

This report presents measurements performed between four European high performance computing (HPC) centers to investigate the possibilities and challenges of European Meta Computing using broadband connections. A high speed, low latency pilot network consisting of Ethernet, FDDI, Datex-M and ATM sections was set up between workstation clusters at the computing centers.

The dynamic load balancing environment HiCon [Beck95e] was installed on the clusters and several complex, parallelized applications - image recognition, finite element analysis and database processing - were executed and observed. Additionally, several of these applications where executed concurrently in the system. The load balancing environment matched the trade-off between resource exploitation and communication overhead according to the system and network behavior. Performance was compared to the nowadays available Internet connection.

The trials were supported by the European Commission among a set of different distributed computing trials within the E=MC^2 project [Horn94], [EMC95c]. The results strengthen common promising expectations and show several interesting challenges, limitations and guidelines for European Meta Computing - resource sharing between distant high performance computing centers by high speed networks and flexible load distribution services.

Table of Contents



Fine Grained Workload Distribution Across Workstation Clusters of European Computing Centers Coupled by Broadband Networks

Wolfgang Becker

Faculty Report No. 1995 / 9

Institute of Parallel and Distributed High-Performance Systems (IPVR)
University of Stuttgart, Breitwiesenstr. 20-22, D-70565 Stuttgart, Germany
Phone: +49 711 7816 433, Fax: +49 711 7816 424
wbecker@informatik.uni-stuttgart.de
http://www.informatik.uni-stuttgart.de/ipvr/ipvr.html

This work was partially supported by the European Commission, Project B2010 TEN-IBC E=MC^2

1 Introduction - The TEN-IBC E=MC^2 Project

The TEN-IBC (Trans European Networks - Integrated Broadband Communication) program [Roy94], [TEN94a], [TEN94b] shall provide detailed guidelines that help identifying the missing elements in the harmonious development and functioning of the infrastructure for the provision of the Trans-European networks - Integrated Broadband Communications. The guidelines shall set the network development approach and yield sound and economic investments for implementing the networks. TEN-IBC defines generic wide-area broadband services and identifies favorable conditions for the deployment of the broadband communications.

TEN-IBC investigates and evaluates different application types for a European high speed network. The results will be disseminated to related interest groups and user communities. Application domains investigated within TEN-IBC are cooperative work, distributed information services and parallel / distributed high performance computing (E=MC^2). TEN-IBC is focussed on application level rather than on network level.

The E=MC^2 (European Meta Computing Utilizing Integrated Broadband Communications) project within the TEN-IBC program [EMC94a/b], [EMC95a/b/c], [Horn94] addresses broadband interconnection of High Performance Computing Centres for remote and distributed use of super computers to meet the challenge of major applications and optimize their use across dispersed research teams. Between autumn 1993 and summer 1996 it identified user interests and evaluates the impact of Europe-wide broadband network availability on the use of supercomputers and computing clusters by research agencies and commercial users.

High Performance Computing (HPC) is severely limited by network bandwidth in the degree to which the resources available at HPC centres throughout Europe can be harnessed. The Rubbia report identified the need for high bandwidth connectivity as a key factor constraining the economic benefits of co-operative uses of HPC. E=MC^2 set up a trans-national ATM based broadband network and exploits it for applications which demand high network performance for their scientific or commercial success, or even for them to be feasible at all.

The project involves several HPC centres, a telecommunications value added services company and a commercial HPC manufacturer (section 5).

E=MC^2 runs several trials, each concentrating on a different aspect of network requirements and reflecting applications of strong interest in the HPC community. The trials concentrate on 1) coupled computing in modelling and simulation using an atmospheric and oceanographic simulation as its application, 2) distributed supercomputing and workstation clustering between the centres and 3) remote submission and execution of applications to investigate the potential for commercial brokerage of HPC services. This report concentrates on the second area. For more detailed information about the whole project and the results obtained so far, we refer to [EMC94a], [EMC95b].

The TEN-IBC E=MC^2 issues are a hot research topic and will become a commercially promising technology in the next years. The concepts for distributed parallel computing mainly come from distributed operating system research [Tane85], [Gosc91] and database processing [Rahm93]. Recent network technologies enable local and wide area networking for HPC with very high performance [Hand91], [Lin94], [McCa94], [Wolm94], [Stru95], [Tolm95], [Pozz95]. Based on this, network computing environment prototypes are being developed [Cap93], [Blum94]. Recently, people even start working on parallelization and distribution of relevant industrial codes [Mech95], [Buba94], [Mier95], [Colb95].

2 Purpose and Rationale of the Trials

The main issues and purposes of E=MC^2 can be summarized as follows:

- Investigation of technical feasibility of wide area distributing compute load across nation boundaries.
- Identification of software support requirements for proper distribution of large compute work load.
- Observation of bandwidth requirements / bandwidth utilization for distributed execution of different parallel application classes and multiuser load. Monitoring of application behavior and network utilization.
- Review potential benefits from trans European high speed networks for commercial and scientific HPC. Obtain user requirements and user satisfaction to evaluate marketability.
- Identify economically important and promising HPC application domains that can utilize European high speed networks, showing completely different profiles and challenges than the more common and more understood domain of video/voice transmission and cooperative work.

The project is not expected to yield exciting discoveries, but to provide a quantitative base for future planning and activities in this area which can be of important commercial benefit in the near future.

The E=MC^2 trials are planned and set up according to following principles:

- Take existing applications. Adapt them as necessary for wide area distribution and heterogeneous network capacities and processing nodes.
- Use IP over ATM rather than native ATM protocols, because all existing parallel / distributed applications are based on proprietary or on IP mechanisms.
- Couple several distant High Performance Computing (HPC) centers. These centers are currently connected by narrowband ethernet, and were connected by high speed ATM during the broadband trials.
- Choose representatives for relevant and typical application classes:
  - European weather forecast simulation coupled with oceanographic simulation.
  - Parallel grid based numerical simulations.
  - Workstation cluster load distribution services with different sample application loads.
  - Client-server structured parallel applications with automatic load balancing support.
- Network monitoring characterizes the application behavior and network utilization patterns.

The main topics for investigation in the load balancing trials will be the following:

- Will the introduction of broadband connections enable a shared, increased utilization of the huge computing resources distributed among the various high performance computing centers within Europe?
- In which degree can these distributed resources be effectively exploited? The granularity and flexibility depends on the availability and capacity of trans-European network connections, on the capabilities of underlying load balancing services and runtime environments and also on the applications' flexibility with regard to migration and task decomposition across heterogeneous structured networks of dissimilar computer architectures.

  The coarsest level of load distribution is the submission of whole large batch jobs to suitable computing centers. This does not require sophisticated software support nor network capabilities, except that all services must be available at each site and input data must be copied to the executing site.

  A finer distribution granularity is the assignment / migration of processes throughout the network. Due to unpredictable arrival rates and processing demands in interactive operation this should be done dynamically and hence requires operating system support for transparent file access, service availability and process migration. On-line network connections with sufficient bandwidth are required.

  The finest distribution level assigns even small subtasks of large complex parallel applications to processors within and between computing clusters to increase the available parallel processing power. This approach additionally requires advanced system services and powerful point to point communication facilities with low latency for synchronization / cooperation within parallel applications and remote access to fine grained data items.

  Overall, the question is whether it is feasible to use the European computing centers as cooperating clusters or even as a large parallel metacomputer.
- Under which circumstances is it worthwhile to distribute and shift load mixes of several concurrent large parallel applications among computing centres to improve performance by load balancing? What connectivity is required to enable a significant and effective reduction of load skews between distant computing centers, provided a suitable load balancing service is available?
- Under which circumstances is it worthwhile to distribute one single large parallel application among computing centres to improve performance by load balancing? How far can coupled computing centers be viewed as one large resource pool, increasing the amount of available resources for HPC?
- What technological challenges and economical constraints limit automatic load distribution for exploitation of the European computing resources? By adapting and upgrading the load balancing environment within these trials and porting different application types of common interest, we try to identify the most critical issues to be improved in nowadays operating systems and load distribution environments.

Which performance criteria apply to the investigated objects?

- The first object type, workstation clusters at computing centers, are observed as to how much of the available processing power can be fruitfully utilized within the clusters and globally.
- The second object type is the network itself. The trials will focus on the average latency of short messages (packets) that are exchanged to synchronize within parallel applications and ship control information throughout the distributed runtime environment, and on the average latency for mixed short and long messages that are sent for data communication and data movement purposes. The overall utilized network bandwidth is not of primary interest in the trials. The measurements obtained from the narrowband connection will be compared to the ATM results.
- The third object of investigation is the load balancing service. The performance evaluation criteria comprise the achieved resource utilization, the average achieved proportion of computing time to communication time per task execution and the throughput gain by coupling several unequally loaded remote clusters. An important issue is the effectiveness of load balancing decisions, i.e. how far the trade-off between exploitation of idle remote CPU cycles and increased data communication overhead can be covered.
- Finally the applications are subject to evaluation. The users' performance criterion is simply the elapsed time of a whole application. The trials evaluate the potential of three different application types for automatic distribution across nation boundaries by load balancing services. The potential depends on the applications' work load profiles (task granularity and potential parallelism), their communication intensity and also on the complexity to properly distribute their tasks across the nodes.

Overall, the different objects and their evaluation criteria capture the interests of users, application developers, operating system developers, computing center operators and network providers.

Following technical challenges were to be considered for the trials:

- Current parallel and distributed computing on workstation clusters suffers from insufficient concepts for naming, security, vendor independence and availability, which make remote processing, global file access, communication and management difficult:

  There is no useful or complete hierarchical naming scheme for countries, computing centers, hosts or network interfaces. Default directory paths for executables and users' homes are different and user names or identifiers often must be different. Illegal data access and resource usage across the network is prohibited by inflexible 'fire wall' routers at the computing centers that render remote execution and communication more difficult. Heterogeneous computer architectures and versions of the UNIX operating system differ in definition files and library functionality; They use different byte orders and byte alignments for data structure representation. Differing executables and data file formats must be managed additionally to the version management of programs, configurations and data across the network and complex scripts are necessary to manage and distribute the versions across the network. Each time, as network lines or participating hosts become unavailable, configuration changes are necessary.
- On the application and load balancing level, the IPVR provided an environment including several applications that turned out to be suitable for low bandwidth as well as for broadband trials. For other existing relevant applications, proper parallelization, restructuring and tuning as well as enhancing their portability and flexibility requires huge efforts that can hardly be afforded within this project.
- The proper exploitation of distributed system resources across the network requires not only equalization of processors' load, but also consideration of the data communication within parallel applications and the cost for accessing non-local common data items. The load balancing structure can be clustered according to the geographic distribution of the workstation clusters.
- The timely network availability and reliability necessary to perform the trials was not clear. Missing experience with the new technology at the manufacturers, network providers and administrators could also infer serious problems with regard to network management and costs.

3 Load Balancing Environment

Clusters of high performance workstations not only provide individuals with personal computing resources, but can be exploited as parallel computing systems. The workstations within clusters are usually not used by their owners all the time but are idle for about 90% of the time. Hence, for batch and interactive processing demands of the users the resources of idle workstations within the cluster can be exploited. Further workstation clusters can be used as a parallel system for large compute intensive applications. To achieve a reasonable exploitation of the available resources and leave programmers and users almost unaware of distribution aspects, dynamic load balancing is required to schedule arriving tasks within the cluster towards optimal throughput.

Different computing centers supply clusters of workstations that meet a relevant, growing portion of the scientific and commercial processing demands. These workstation clusters can also be connected to one huge processing array for large computations or yield an optimal utilization of system resources by different users and applications entering the system from local or remote sites. However, to exploit this metacomputing ressource for the European research and commercial community, suitable broadband connections between European computing centers as well as automatic dynamic load balancing facilities operating within and between clusters are basic requirements.

In contrast to the remote execution trials within $E=MC^2$, the load balancing trials do not focus on a general purpose environment for mixed load by anonymous processes, but on load balancing facilities for increasing the overall system throughput of several big, heterogeneous applications typical for commercial and scientific production environments. Here, the proper exploitation of distributed system resources across the network additionally requires consideration of the data communication within parallel applications and the cost for accessing non local, common data items. Further, to efficiently equalize load for large, important applications it is worthwhile to provide load profile estimations of the applications‘ tasks and data to the load balancing facility. Unlike other common load balancing and remote execution environments, the HiCon load balancing environment (see below) can deal with parallelized applications with heavily communicating tasks, synchronization relationships and cooperation on common data structures. Exploitation of distributed system resources across the network requires, besides full utilization of the CPUs, also consideration of data affinity and data communication overhead when spawning tasks that operate close together.

A full implementation of the dynamic, workstation based load balancing environment HiCon, developed by the IPVR was employed. In this environment load balancing is expected to assign work load in a way that maximizes overall throughput. The execution environment offers advanced, flexible and adaptive dynamic load balancing for compute and communication intensive applications on parallel shared nothing architectures like workstation clusters. Although application independent, load balancing is able to exploit various resource information units for decision making, like CPU run queue lengths, task queue lengths, location of data or copies and estimations about task profiles. It can operate different complex, reactive and predictive strategies. Further, load balancing is able to dynamically adapt its decision parameters to the current application and system state. Different clustering structures between centralized and completely distributed load balancing structures can be configured. For the trials it is essential that the load balancing system is scalable and flexible enough to span across clusters of heterogeneous workstations of high performance computing centers in Europe. Task and communication granularity as well as communication patterns can be observed in different granularity. Summarizing, the HiCon system provides following features:

- Dynamic task assignment, application independent, adaptive. For heterogeneous workstation clusters, for large, parallel applications, for multiple concurrent applications.
- Central & local task queues
- Centralized per cluster, decentralized between clusters
- Decision cost model: Minimum estimated task response time: wait time in local server queue + compute time on node + delays for data communication
- Decisions base on measurements of nodes’ load, data distribution, data exchange cost (including network bandwidth and latency) and pre-estimations of task size, data access pattern
- Applications observed under load balancing: Task parallelized, client - server structured, fine granular communicating tasks within applications, cooperation by accessing virtual shared data (runtime system support)

Applications obey the client - server processing scheme, they are functionally decomposed, and one client per application which controls the application flow, issues tasks to be worked off by some instances of a server class. Of each server class multiple instances can be distributed among the processors. There is no explicit communication between servers; instead, they cooperate by using global data structures, actually automatically distributed and replicated among the computing nodes. Parallelism within applications is achieved by asynchronous server class calls.

During the trials we will not look closer to load balancing policies, load balancing overhead or adaptability of dynamic load balancing strategies to network load patterns. We just employ the most sophisticated and suitably tuned and load balancing structure and policy respectively.

4 The Trials: Application Scenarios

A set of applications was chosen, which is representative for the behavior of large important applications in the areas of numerical simulations, image processing and commercial data processing. The applications are parallelized and nevertheless can be run concurrently as homogeneous or heterogeneous load mix on arbitrary workstation networks within the load balancing environment. The following gives a brief characterization of the applications used within the load balancing trials.

Parallel Finite Element Analysis

The finite element method is mainly used in the mechanics and thermodynamics to investigate the behavior of constructions and volumes under the influence of forces and temperatures by numerical simulation. The objects to be investigated are decomposed into a large number of small elements, between which the physical interactions are calculated by approximation of the respective partial differential equations. This element calculation process yields a large global linear equation system, which must be solved to get the resulting displacements, stress or temperature of the elements. The figures show an example result of a block under stress, and the execution scheme of a parallel finite element analysis.

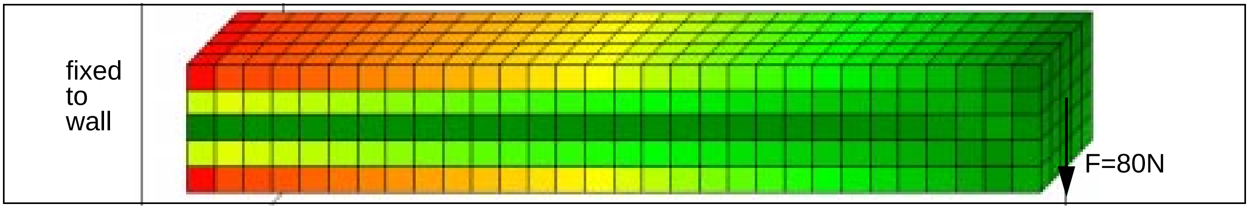


Figure 1: Result example of a finite element calculation.

The algorithm is a major representative for computational load by large scientific simulations. The element calculation is parallelized by element number ranges, a global stiffness matrix is established using a distributed storage format for sparse matrices. The following conjugate gradient solver contains three parallelized sections per iteration. Overall, the application type shows regular task parallelism, stable task sizes and stable data reference patterns. All phases of the computation require heavy data communication, because the tasks operate on common data elements.

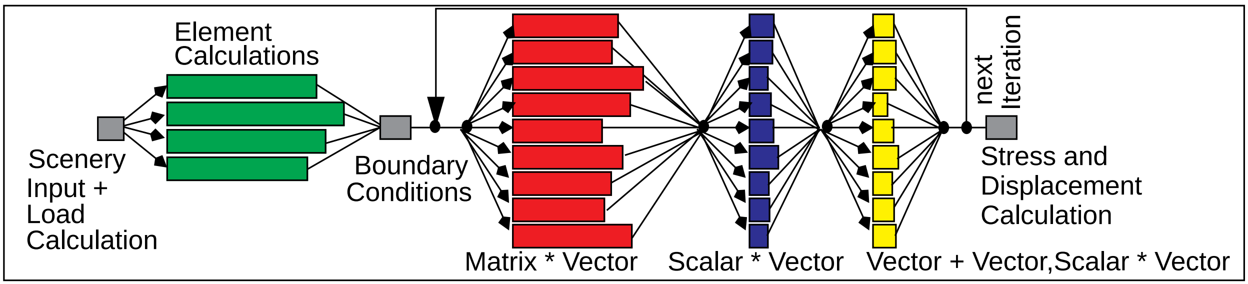


Figure 2: Execution scheme of a parallel finite element analysis.

Parallel Image Processing

Image segmentation converts a given pixel image into a set of homogeneous areas. This process usually is the first step of a complete image recognition. Each area shall consist of similar colored dots with few exceptions. The implemented algorithm takes four steps: Starting from an initial partitioning, the square merge phase tries to combine neighbored squares as long as the resulting squares still give a homogeneous area. Parallel to this process square split operations refine squares that could not be merged, until each square represents a homogeneous area. These both processes are blockwise parallelized and work concurrently with irregular parallelism on the image. The next step, polygon merge, merges as many as possible neighbored squares to arbitrary polygons. The last step, boundary extraction, yields the edges surrounding the polygons.

The split and merge operations are mostly very fine grained. Some can be bundled into larger calls for picture segments. Overall the number of tasks and their granularity depends heavily on the respective structure of the image and influences the parallelism that can be fruitfully exploited. During the polygon merge phase it is no longer possible to target the calls to disjoint image partitions, because the polygons have arbitrary shapes. Depending on the image structure this causes intensive data communications and limits the useful parallelism. This observation also holds for the following calculation of the polygon boundaries.

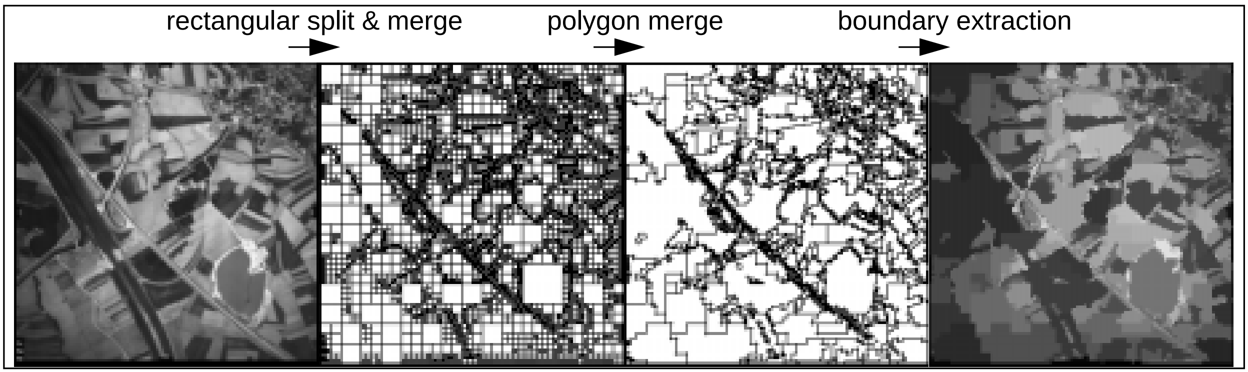


Figure 3: Steps of the image segmentation algorithm.

This class of algorithms, typical for image processing algorithms, shows several computation stages, unstructured parallelism, differing task sizes and frequent data communication; load profiles depend heavily on the actual structure of the image.

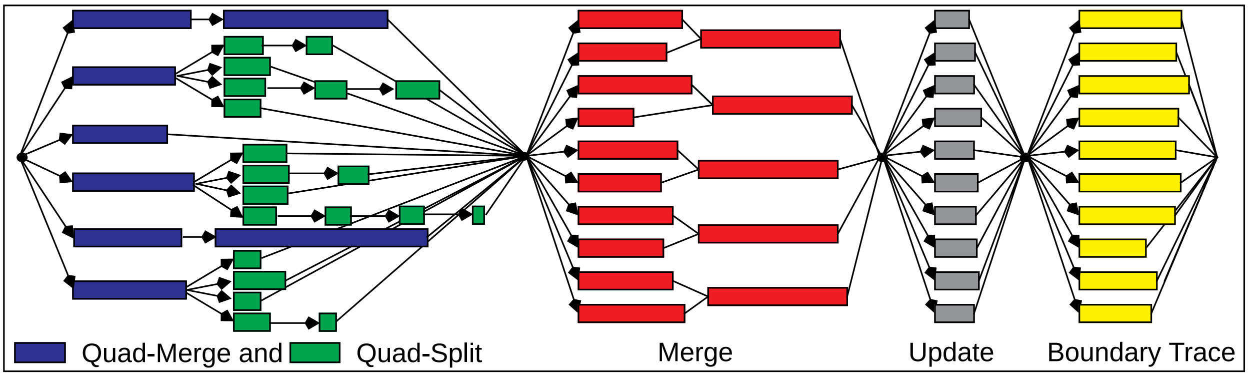


Figure 4: Execution scheme of a parallel image recognition.

Parallel Relational Database Processing

In commercial data processing, a large portion of the efforts consist of retrieval, combination, filtering and modification of data items within a huge set of semantically correlated data. Relational database management systems convert complex, descriptive operations into a set of simple basic executions.

Here, complex query graphs, consisting of scans, projections, joins and loads are performed exploiting functional and data parallelism on key range partitioned data, yielding execution profiles characteristic for commercial data processing. The figure shows a sample query graph along with a possible parallel execution scheme. In this application type, task sizes and data reference patterns can be pre-estimated.

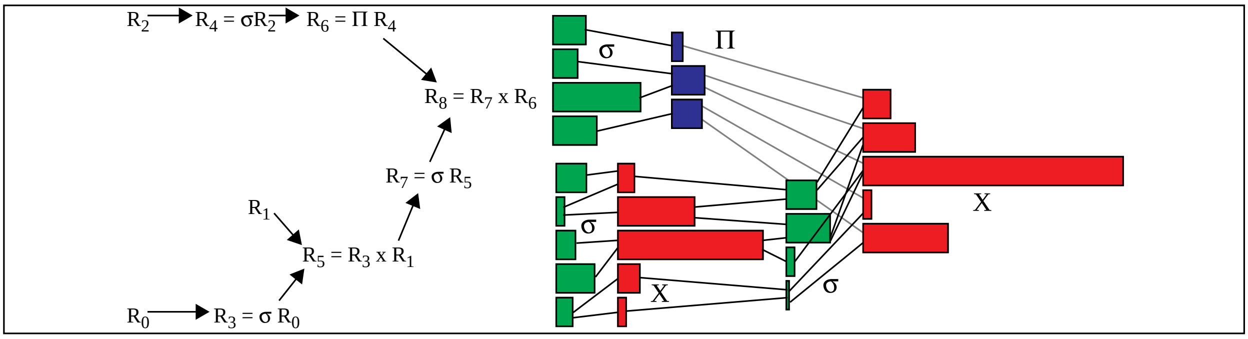

$R_2 \rightarrow R_4 = \sigma R_2 \rightarrow R_6 = \Pi R_4$
$R_8 = R_7 \times R_6$
$R_7 = \sigma R_5$
R_1
$R_5 = R_3 \times R_1$
$R_0 \rightarrow R_3 = \sigma R_0$
σ
Π
X
σ
X
σ

Figure 5: Algebraic description and parallel execution scheme of a sample query.

Network Requirements

The feasibility of load equalization across computing centers depends on the capabilities of the underlying network connections. The requirements for the trials can be defined by the message traffic profile imposed by the load balancing environment. Within one workstation cluster clients and servers exchange call and result messages which are routed via the centralized load balancing component (point-to-point messages with a size of about 100 bytes and a frequency of 0.05 to 50 messages per second, depending on the task granularity and the parallelism within the applications). These messages are almost latency bound.

Between system partitions the same type of messages for task migrations and remote result returns occur between the load balancing components. Their frequency depends on the load skew tolerance of the inter-cluster load balancing policy (rough equalization yields low message traffic). Further the load balancing components interchange system load information messages of 50 byte size in adjustable periods (usually seconds).

Data communications between servers involve short control messages (about 50 bytes) to the load balancing component or probable data owner and longer messages for the actual data transfer, which occur directly between the current owner of the data and the requester node. The size of the data messages is application dependent (usually between 100 and 10.000 bytes), the frequency depends on the data requirements of the applications‘ tasks and also heavily on the data affinity of the servers to the requested data, which can be increased by advanced load balancing strategies. These messages are almost bandwidth bound. Further, there are short control messages for data copy invalidations, and for data transfer across clusters additional forwarding messages between the concerned load balancing components are necessary.

Overall, the message traffic in the load balancing environment consists, besides control, task and result messages, mainly of data communication messages between server processes. There are situations, showing thousands of data communication messages per second that easily congest low bandwidth networks. Messages transferring large data items demand high bandwidth, shorter ones are latency bound. Although message latency can be hidden partially by multiple servers per node and different concurrent applications, the progress of parallel applications heavily depends on fast message delivery.

The load balancing environment uses the UDP message passing protocol (TCP/IP family) rather than ATM specific protocols for portability reasons.

5 Computing Center, Network and Application Configurations

E=MC^2 couples several European HPC centers to exploit the aggregated computing power consisting of supercomputers, parallel systems and workstation clusters. Figure 6 shows the project partners, Figure 7 the network connectivity that was set up by the project.

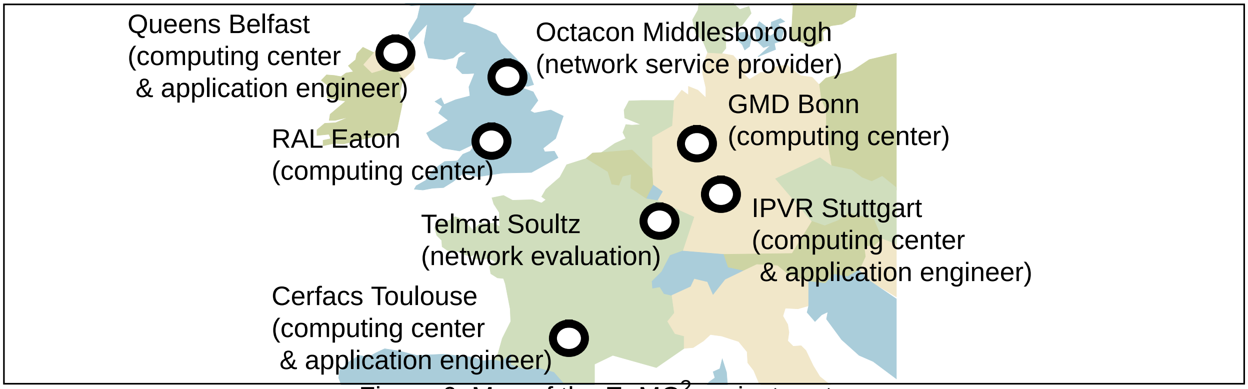


Figure 6: Map of the E=MC^2 project partners.

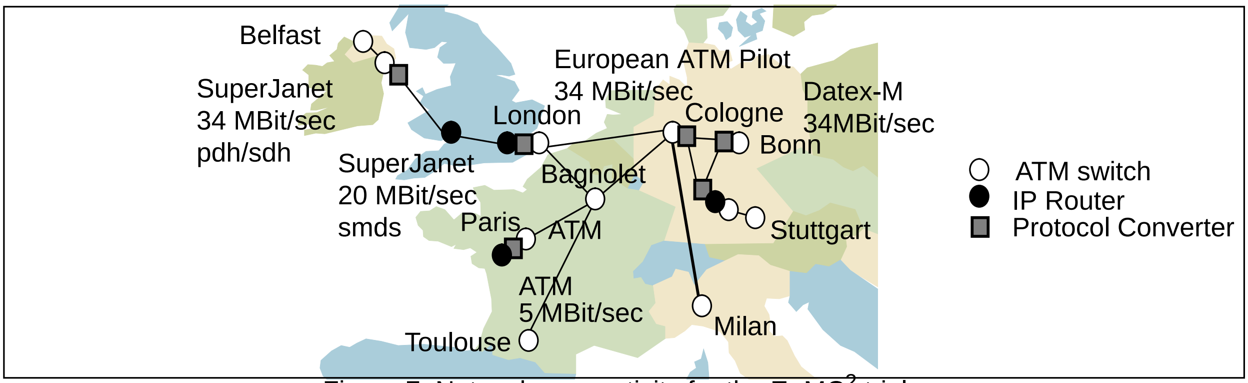


Figure 7: Network connectivity for the E=MC^2 trials.

Figure 8 illustrates the general scenario for the load balancing trials: the computing centers participating in the projects each provide workstation clusters. A central load balancing component was used for simplicity and to increase load balancing accuracy by global overview.

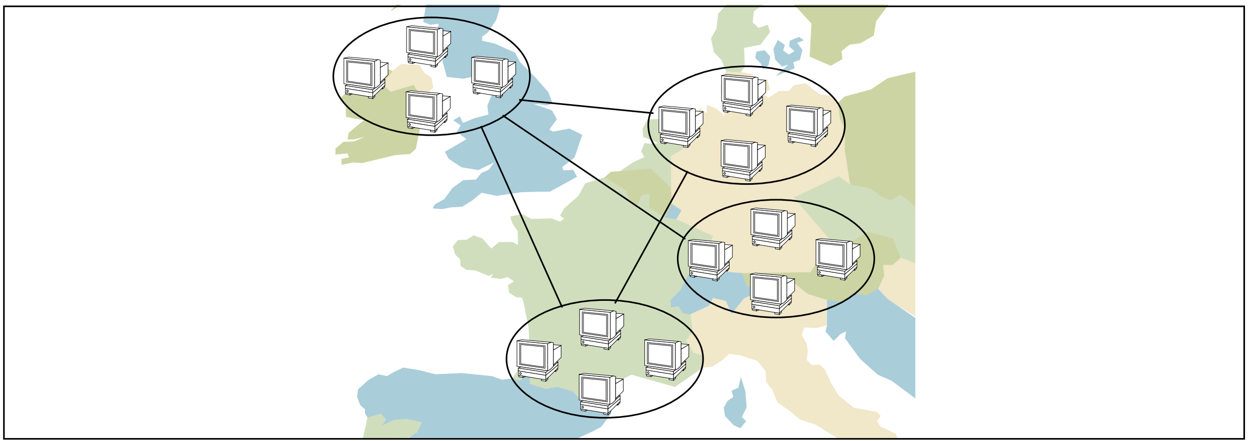
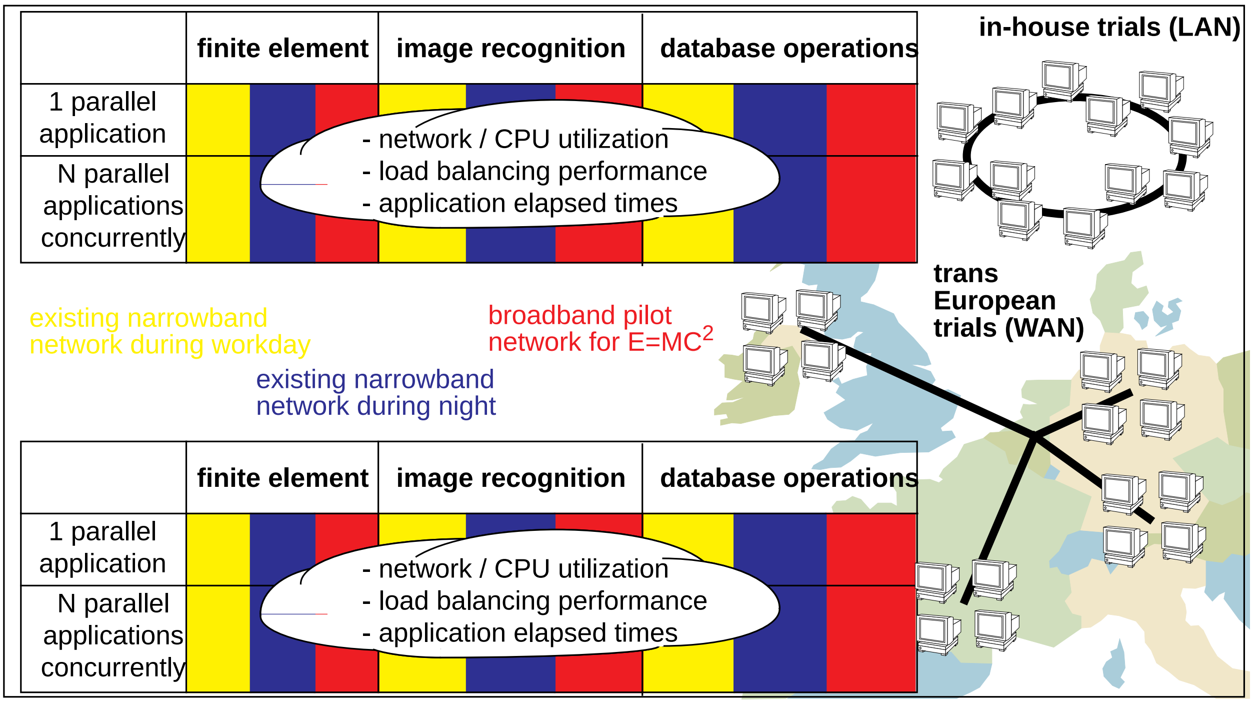


Figure 8: Clustered environment for wide area networked parallel computing.

The general network and application topology for the load balancing trials consists of all four HPC centers CERFACS, GMD, IPVR and Queens. The environment and the applications can run across multiple platforms. Here System V and BSD UNIX derivatives from IBM, Sun and Silicon Graphics are involved. Different network capacities were investigated, each of them was evaluated for different load profiles, i.e. application scenarios. Possible network configurations, as shown in the figure, are locally Ethernet connected workstations, locally ATM connected workstations, trans European narrowband connected workstation clusters and trans European high speed connected workstation clusters.



	finite element	image recognition	database operations
1 parallel application	🟨🟦🟥	🟨🟦🟥	🟨🟦🟥
N parallel applications concurrently	🟨🟦🟥	🟨🟦🟥	🟨🟦🟥

	finite element	image recognition	database operations
1 parallel application	🟨🟦🟥	🟨🟦🟥	🟨🟦🟥
N parallel applications concurrently	🟨🟦🟥	🟨🟦🟥	🟨🟦🟥

- network / CPU utilization
- load balancing performance
- application elapsed times
- network / CPU utilization
- load balancing performance
- application elapsed times
in-house trials (LAN)
trans European trials (WAN)
existing narrowband network during workday
broadband pilot network for E=MC^2
existing narrowband network during night

Figure 9: General arrangement of network and applications for the load balancing trials.

Of course, due to time limitations and unavailability / unreliability of the broadband connections not all cells in the tables above could be filled by realizing corresponding measurement scenarios. Instead, in the following the successfully realized trials are presented.

6 Measurement Results

Although the E=MC^2 project performed a variety of different measurements, it is impossible to investigate network and resource utilization and the application performance on a continuous line of network speed.

In the following, the trials are presented for the different topologies that were employed. Each topology is described in three aspects:

- the network routing,
- the network performance and
- the application performance.

For each topology we tried to compare network lines of differing capacity, namely the existing Internet and the broadband pilot lines set up especially for the project.

Network Routing

The routing of message packets across the existing narrowband network within Europe is rather complicated and not straightforward. The figures show typical routing lines between the participating computing centers within E=MC^2. Therefore, the IP *traceroute* utility measures the average round trip delay for sending short UDP packets (40 bytes) across the network and traces these delays for each router along the path. Hence, it provides the whole route which the packets take and the time spent on each section. The latencies vary extremely, depending on the traffic profiles, which change during day and week time. Hence, the routing changes dynamically and may also contribute to latency variations. For example, between Stuttgart and Toulouse, latencies in the range of 200 ms to 600 ms can be observed. The same degree of variation holds for the network performance measurements.

Not only the far distance lines, but also the large number of hops within the switching stations, i.e. network provider exchanges or computing centers, add significant delays to the message delivery time. Hence, the broadband pilot connections established for the E=MC^2 trials not only largely improve the throughput, but also reduce the latencies significantly. Further, the routers and switches are also the main reason for the packet loss rates, because they drop packets in congestion situations.

Within the UK the research centers are already connected by SuperJanet broadband lines for the usual Internet services. Instead, France and Germany are equipped with region wide broadband networks only, like the Belwue in Baden-Württemberg.

Network Performance

The main focus of the E=MC^2 research is on the application level, where the application performance of course depends heavily on the performance and characteristics of the underlying networks. So it is essential to judge the network performance. The raw network level throughput is the most commonly used measure for network performance. This was also observed in depth [EMC95a], but the E=MC^2 consortium identified further characteristics as important and tightly correlated to performance on application level:

- Distributed HPC applications usually exchange data and synchronization messages between the participating processes. This yields no continuous data flow and hence no effective network utilization. However, it requires very high bandwidth during certain phases within an application execution.
- Most applications are parallelized in a way, that the progress of the participating processes depends on the arrival of certain data or synchronization messages. Therefore, the elapsed time for a single, possibly short message is a limiting factor for the performance of distributed applications. This is why the network latency and the throughput for single / synchronous messages is an important network characteristic. Synchronous messages are used frequently, if the underlying delivery service is unreliable, or if the sender requires that the receiver really has dispatched the message, possibly just to ensure that the partner has also reached a certain point within the common computation.
- The Internet protocol (IP) provides unreliable packet delivery. Packets can be lost on their way arbitrarily, due to falsification of the bit contents or due to congestion of the input / output queues of the intermediate routers and switches. There is no notification mechanism that tells about lost packets. Hence, the protocols basing on IP usually optimistically send several message packets across the network and besides listen for acknowledge packets. After some time or if a certain amount of data has been sent, and no acknowledge returned so far, the sender protocol simply retransmits the packets. So,

the average packet loss rate also significantly influences the network throughput visible to the application, because lost packets entail large delays for the time-outs and additional traffic on the network due to retransmissions.

The network performance graphs present the relevant performance characteristics of the networks used by the trials. It should be mentioned, that the performance on the existing narrowband wide area networks show huge derivations in all aspects. So the numbers in the graphs merely reflect typical average behavior.

Application Level Measurement Results

As depicted in section 5, we tried to execute and observe as much as possible different scenarios on the different network topologies. The results on application level give elapsed time of a whole parallel application's execution or elapsed time of a whole mix of concurrently executing parallel applications. This serves as a judgement of response time and of system throughput as well.

6.1 In-house Broadband Trials

Within the IPVR, some trials were performed to compare the performance of the Ethernet CSMA network with up to 10 Mbit/sec bandwidth to the ATM switched network providing 100 Mbit/sec (TAXI). Up to five hosts were connected by either Ethernet or ATM. Figure 10 gives a performance comparison between the Ethernet and the ATM broadband connection in-house within the IPVR. Latency is about the same, because it is mainly software protocol stack processing and context switching time. This holds for messages up to 3 KByte. The throughput, however, increases dramatically.

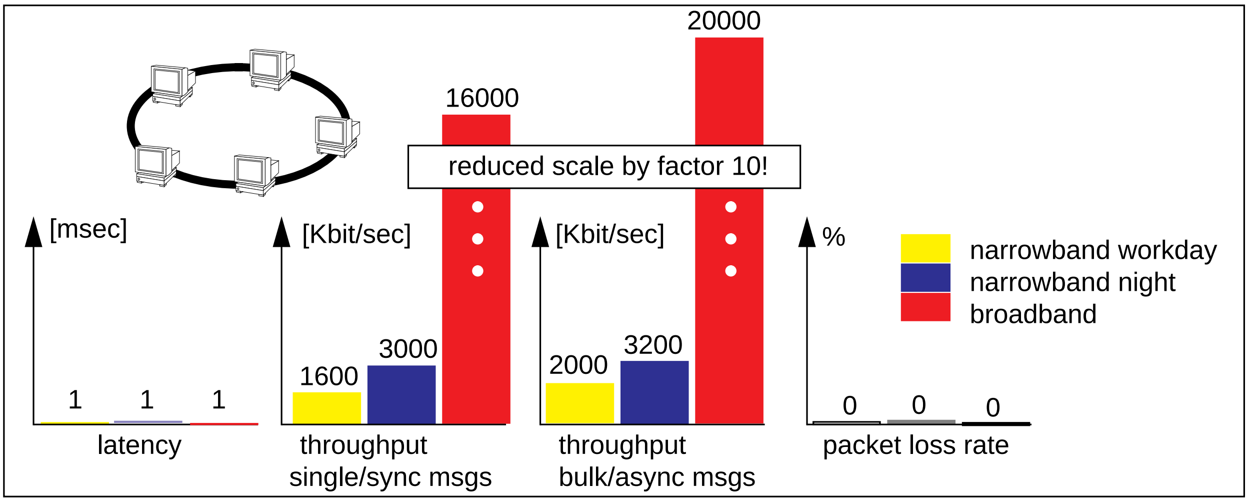


Figure 10: LAN network performance.

Despite the throughput enhancement, the application level measurements showed no significant differences. Load balancing always exploits all available servers. But as this type of distributed computing does not heavily use large data streams, the latency of the LAN is a limiting factor.

For illustration purpose, Figure 11 and Figure 12 observe one of the five participating hosts during the execution of three concurrent and of nine concurrent parallel finite element analysis applications on ATM LAN. By tracing the ATM cells sent / received by this host and underneath displaying the CPU utilization profile, it shows the intermittently usage pattern of the network, corresponding to the calculation phases of the applications. The network utilization increases as expected, if more applications are executing concurrently. Still, they produce no continuous but rather irregular traffic, demanding very high bandwidth for short phases. Note, that this is only the network load by one of the hosts, hence the message traffic crossing the switch is about five times as high, rising up to 5 Mbit/sec sustained throughput in peak situations.

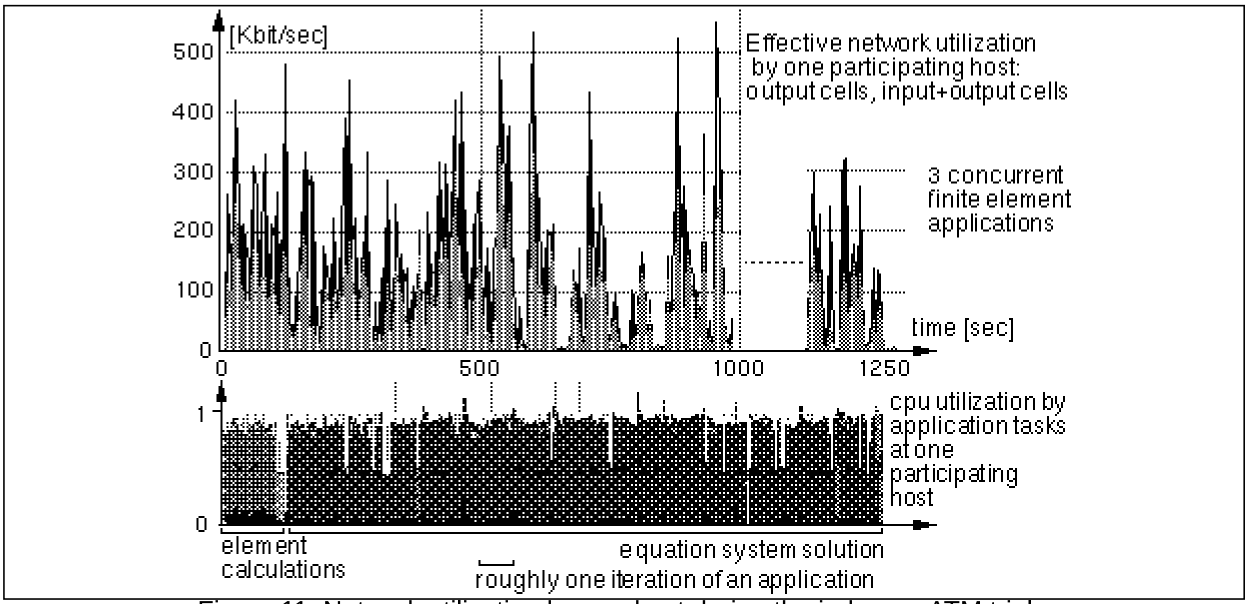


Figure 11: Network utilization by one host during the in-house ATM trials.

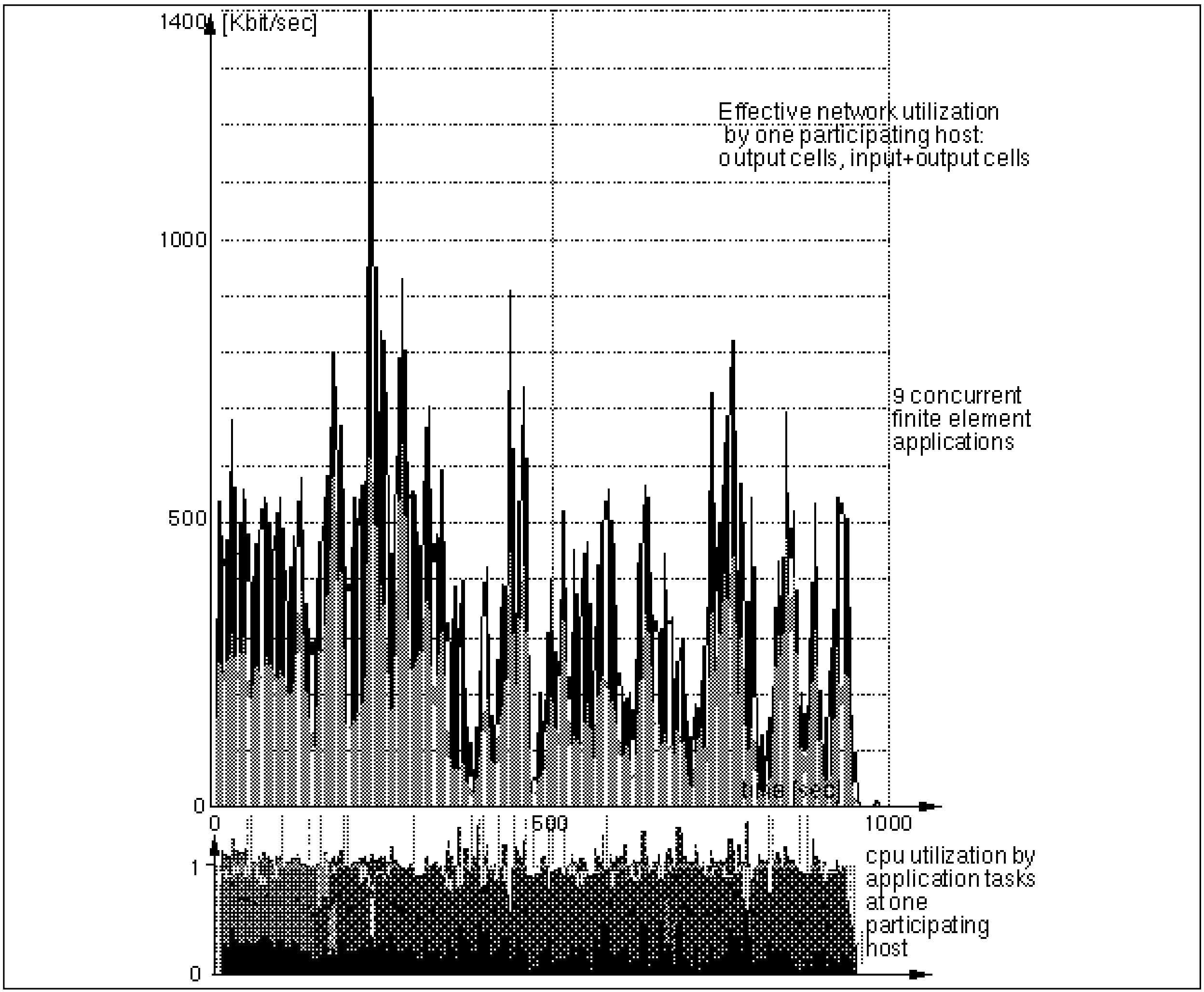
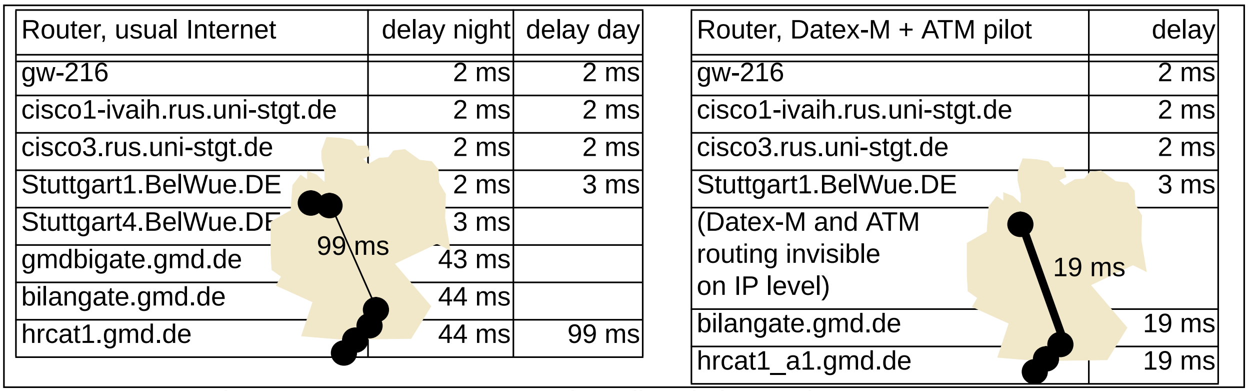


Figure 12: Network utilization by one host during in-house ATM trials with high concurrence.

6.2 National Broadband Trials between Stuttgart and Bonn

Cooperating with the national research laboratory GMD near Bonn, it was possible to set up a Datex-M link with broadband end connections on both sides between GMD and Stuttgart for the E=MC^2 trials. After walking through various problems and challenges which where typical for all the broadband pilots within the project, the line was up for several days with acceptable bandwidth and reliability. First, we look at the routes and message latencies on the existing national Internet and on the national broadband pilot connection established trials between the IPVR in Stuttgart and the GMD laboratories in Bonn (Figure 13). The Stuttgart router converts IP packets into SMDS packets and sends them over an HSSI interface per Datex-M to Cologne. This is based on DQDB technology. In Cologne, the protocol is transformed from DQDB towards ATM. The Bonn router then receives SMDS packets via ATM, extracts the IP packets and packs them up into AAL5 packets. These packets are sent via ATM to the target workstations.



Router, usual Internet	delay night	delay day
gw-216	2 ms	2 ms
cisco1-ivaih.rus.uni-stgt.de	2 ms	2 ms
cisco3.rus.uni-stgt.de	2 ms	2 ms
Stuttgart1.BelWue.DE	2 ms	3 ms
Stuttgart4.BelWue.DE	3 ms	
gmdbigate.gmd.de	43 ms	
bilangate.gmd.de	44 ms	
hrcat1.gmd.de	44 ms	99 ms

Router, Datex-M + ATM pilot	delay
gw-216	2 ms
cisco1-ivaih.rus.uni-stgt.de	2 ms
cisco3.rus.uni-stgt.de	2 ms
Stuttgart1.BelWue.DE	3 ms
(Datex-M and ATM routing invisible on IP level)	
bilangate.gmd.de	19 ms
hrcat1_a1.gmd.de	19 ms

Figure 13: Current Internet vs. broadband pilot routing Stuttgart - Bonn.

A performance comparison between the existing national network and the broadband pilot connection between IPVR Stuttgart and GMD Bonn gives following results:

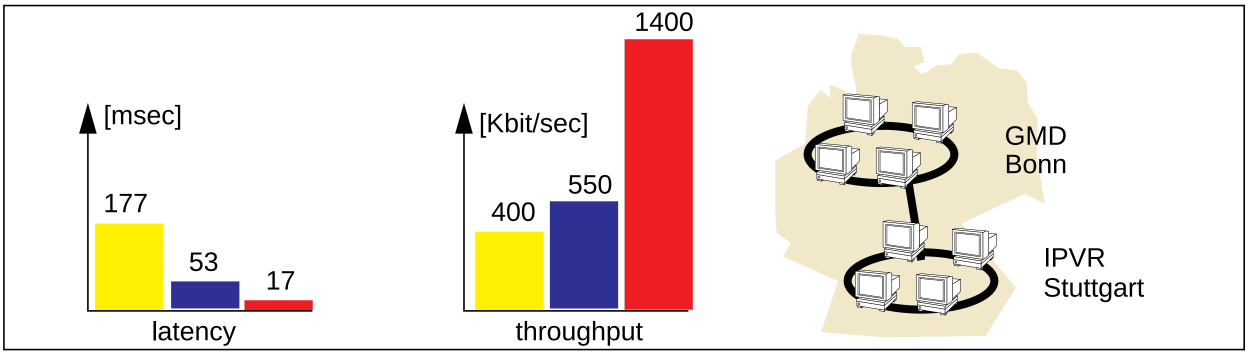


Figure 14: National network performance between Bonn and Stuttgart.

This time, network throughput was measured by observing the UNIX *rcp* command for remote file copying. Because a 1 MB file was copied, the numbers roughly correspond to the bulk / asynchronous values of the measurements in other topologies. The raw network performance of the broadband connection, consisting of mainly Datex-M and smaller ATM / FDDI / Ethernet parts was told to provide a raw throughput of 8 MBit/sec as requested from Telekom. The narrowband connection (WIN) was specified to offer nearly 1 MBit/sec if completely undisturbed. During workdays, however, the network specialists said it could be arbitrarily slow. User level throughput is of course smaller due to *rcp*, IP and ATM protocol overhead. The average message latency could be dramatically reduced by the high speed link.

Based on this broadband connection, the largest variety of different scenarios involving all three application types could be performed. Comparative measurements using the usual Internet connection could be obtained during night and during workdays. The clusters consist of 4 workstations running SunOS (BSD Unix) or Solaris (System V Unix) at each site. At the IPVR, two processor workstations were employed.

First, each application was observed in dedicated mode on the cluster. Next, each of the applications was run several times concurrently. Here, all the 5 application executions were started from one GMD host and appropriately distributed across the network by automatic load balancing. The results are summarized in the following figures.

The parallel finite element analysis was executed by nearly 100% exploiting the IPVR servers and about 75% of the GMD servers. The problem of the executions on the high capacity network is that the waiting times at the end of iterations due to slightly unequal computing times are getting more important as the network communication delays drop.

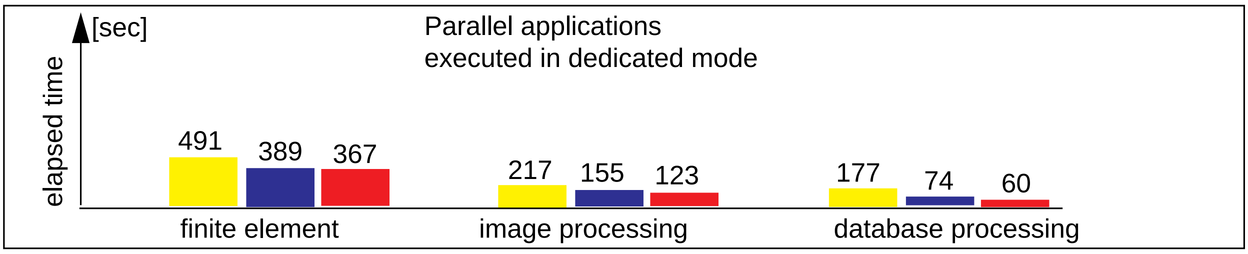


Figure 15: Application level performance of the national trials, dedicated applications.

The parallelized image recognition on the narrowband network was executed 70% at IPVR machines (quad phase). In the merge phase, load balancing exploited an average parallelism degree of 3 using IPVR machines. In the boundary search phase, however, all servers where used. On the broadband network, slightly more parallelism was exploited.

The parallel database operation execution was always executed to 95% at the IPVR, because internal parallelism was not that high and the IPVR machines are significantly faster.

Overall, it must be stated that the rather fine granular distribution of a single parallel application across the wide area network does not show significant improvements in general, indicating that wide area resource sharing is especially worthwhile to increase the computing potential if a local site cannot provide enough resources for the workload it receives, and not to achieve arbitrary speedups by further parallelizing applications.

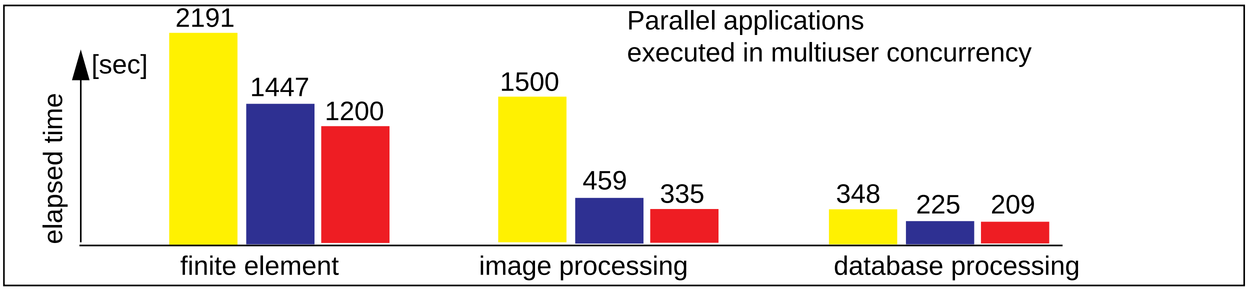


Figure 16: Application level performance of the national trials, multiuser operation.

For the 5 concurrent image recognition computations load balancing decided to exploit all servers for almost all the computation phases. The same holds for the other 2 applications. So, load balancing estimates this network topology as close enough to fully exploit the distant computing clusters as one meta computer. While in following international topologies this feeling can be achieved by the broadband ATM pilot only, here the broadband connection does not principally change the distribution policy but yields largely reduced communication delays and significant better resource utilization. This can successfully be exploited to improve application level throughput.

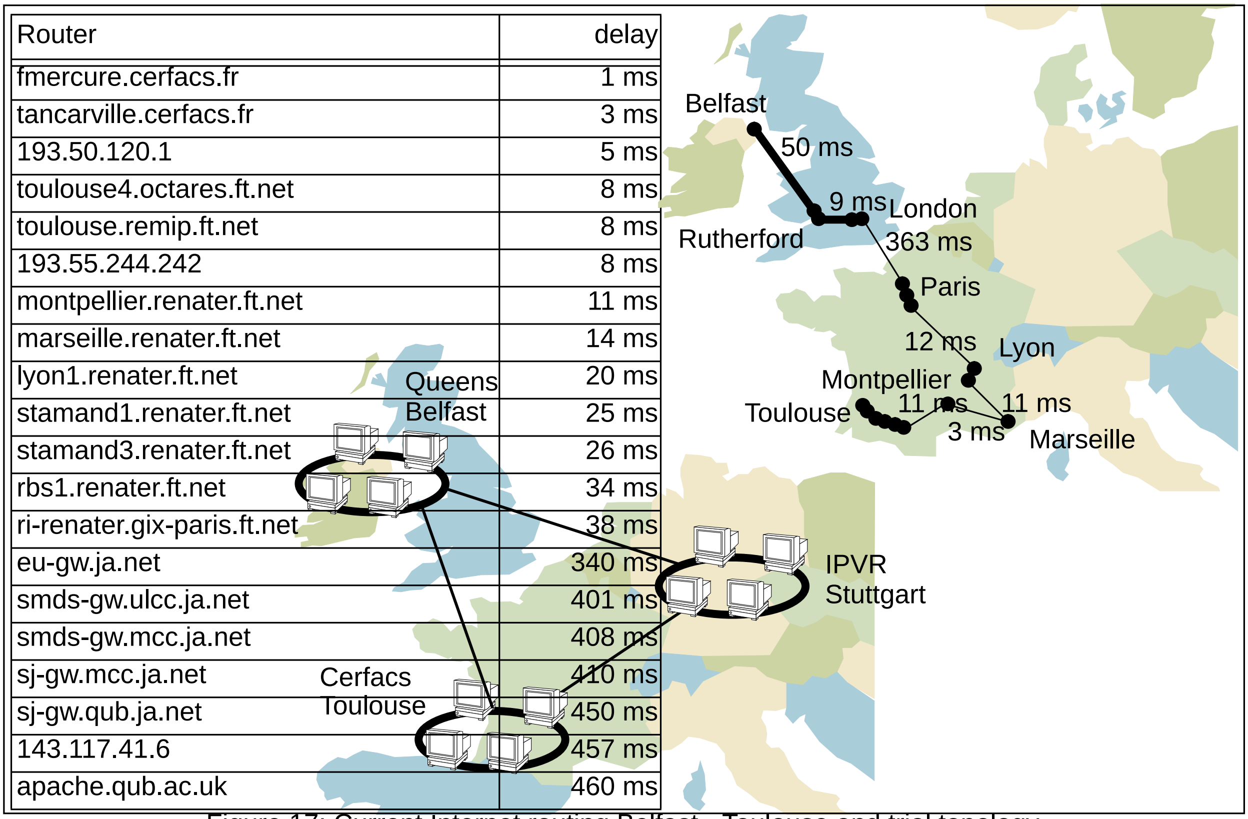
6.3 Trans European Narrowband Trials at Different Time of Day

The performance of the existing narrowband WAN within Europe available for a user varies largely between night and day. During working days the network lines are completely overloaded - [EMC95a] provides extensive measurements. The network routing figures are not shown here, but in following sections which also used broadband lines. Figure 17 just shows a typical Internet routing between Queens University of Belfast in Northern Ireland and CERFACS Toulouse in Southern France.

These trials on the narrowband network were performed to evaluate the sensitivity of the application performance to the underlying network performance. However, it must be mentioned that the network shows a very bad performance both night and day, so that load balancing mostly refused to really exploit this network. Hence, one major conclusion of these measurements is, that the current network really is not capable to enable suitable European meta computing.

The configuration consists of all three distributed computing centers, each of them supplying a workstation cluster as indicated in Figure 17.

The network characteristics for some of the connections that can be obtained for typical distributed HPC applications' communication profiles is shown in Figure 18.



Router	delay
fmercure.cerfacs.fr	1 ms
tancarville.cerfacs.fr	3 ms
193.50.120.1	5 ms
toulouse4.octares.ft.net	8 ms
toulouse.remip.ft.net	8 ms
193.55.244.242	8 ms
montpellier.renater.ft.net	11 ms
marseille.renater.ft.net	14 ms
lyon1.renater.ft.net	20 ms
stamand1.renater.ft.net	25 ms
stamand3.renater.ft.net	26 ms
rbs1.renater.ft.net	34 ms
ri-renater.gix-paris.ft.net	38 ms
eu-gw.ja.net	340 ms
smds-gw.ulcc.ja.net	401 ms
smds-gw.mcc.ja.net	408 ms
sj-gw.mcc.ja.net	410 ms
sj-gw.qub.ja.net	450 ms
143.117.41.6	457 ms
apache.qub.ac.uk	460 ms

Figure 17: Current Internet routing Belfast - Toulouse and trial topology.

In order to motivate load balancing to utilize the distributed resources, five applications of the respective type were started concurrently at the IPVR. Although their initial tasks were performed at the IPVR, the applications soon began to unfold extensive parallelism, computational load and data communication. Load balancing had to decide, whether it was worthwhile to execute some of the tasks at distant sites.

One restriction of the centralized load balancing structure which was used for most of the trials, is that it was not able to simply assign or migrate whole applications among the clusters, because it decides task based. Usually, the effects are similar to distributing whole applications, if the network power differs widely within and between clusters. But sometimes it turns out that load balancing does not use the other clusters, except these where the respective application originated, because when the applications grow in this cluster, distributing some of their tasks to a remote cluster would lead to huge data communication efforts.

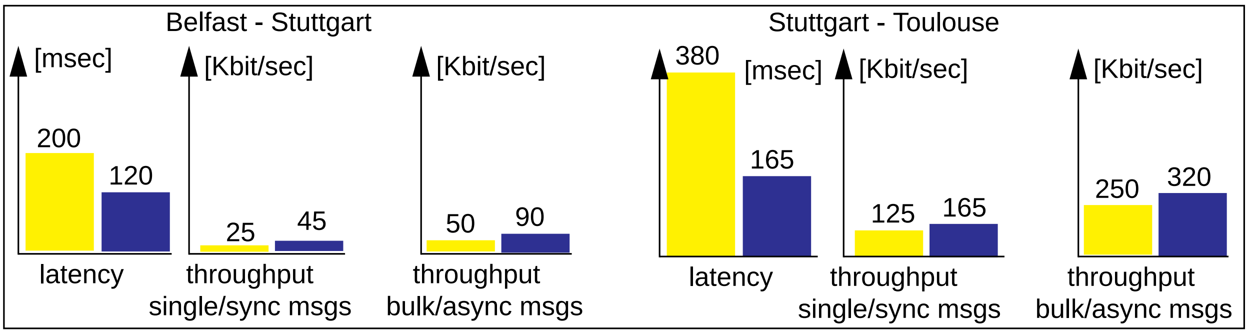


Figure 18: Trans European narrowband performance Belfast - Stuttgart - Toulouse.

Figure 19 shows the elapsed time for the whole application mixes, for each application type and for execution at night and during a workday respectively. To explain the execution time differences, it should be considered, that the CPU resources were quite the same, the applications load was the same, just the underlying network performance changed. Further, load balancing had to decide, how much of the resources it could fruitfully exploit in order to improve the overall performance. Hence, in contrast to statically distributed applications, changing network characteristics yield completely other task assignments and schedules. Finally it should be noted, that load balancing does not primarily try to fully use up the available network bandwidth, but tries to get the workload done as efficiently as possible. This often results in intermittently network utilization.

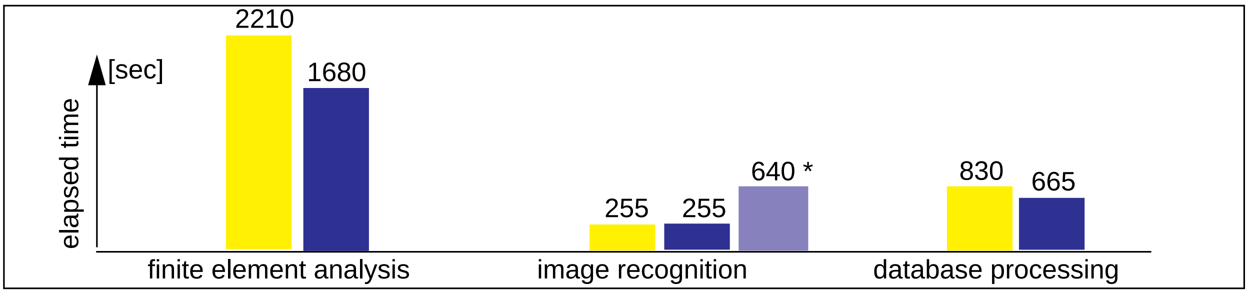


Figure 19: Application level performance of European narrowband trials.

For the finite element analysis applications, load balancing decided to use mostly Stuttgart machines; At night Toulouse and Belfast where utilized slightly more. For the even more fine grained and more communication intensive image processing applications, load balancing decided to use Stuttgart only, regardless of the daytime. If load balancing was forced artificially to use the other clusters also, the application mix took 640 seconds at night. This is just added to ensure, that load balancing decisions are in fact reasonable. For the database processing applications, load balancing fully utilized all clusters in both cases.

These measurements indicate that on existing wide area networks fine grained international meta computing is infeasible, and that HPC application performance is really sensitive to the underlying network capacity.

6.4 Trans European Broadband Trials: Stuttgart - Paris

During the Interop+Networld fair 1994 in Paris, the RUS of the University of Stuttgart set up an ATM / Datex-M broadband connection between Stuttgart and Paris. It was arranged that this connection could be used for E=MC^2 load balancing trials at night.

The message routing and latencies on the trans European broadband pilot connection established during the Interop fair between IPVR Stuttgart and the conference center in Paris are shown below.

Router	delay
gwinf.informatik.uni-stuttgart.de	3 ms
cisco1-ivaih.rus.uni-stuttgart.de	4 ms
cisco3.rus.uni-stuttgart.de	4 ms
Stuttgart1.BelWue.DE	4 ms
(Datex-M routing invisible on IP level)	
153.17.160.31	34 ms
(ATM routing invisible on IP level)	
53.17.170.2	62 ms
paris.rus.uni-stuttgart.de	63 ms

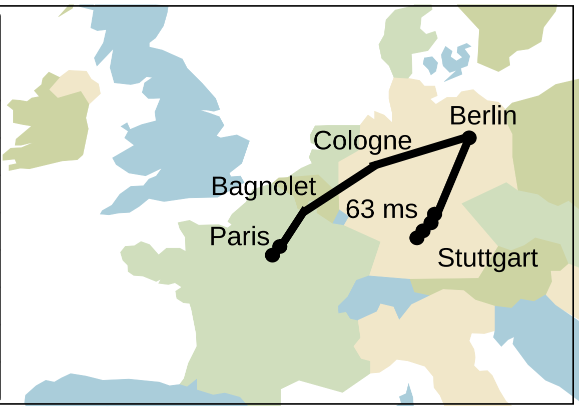


Figure 20: Broadband pilot routing Paris - Stuttgart.

Between Stuttgart and Berlin a 34 Mbit/s Datex-M line was used. Between Paris and Berlin an ATM line was set up with intermediate switches in Bagnolet and Cologne, i.e. the European ATM pilot. At both ends of the ATM line messages where converted from / to Ethernet protocol, which limited the access bandwidth to the broadband connection to 10 Mbit/s. Further, Telekom supplied an OSI level 3 protocol router between DQDB (Datex-M) and HSSI/DXI (ATM) in Berlin, which increased message latencies, because packets were reassembled on IP level.

Overall, several serious technical and organizational restrictions diminished the executable measurements and the strength of the results. These problems, caused by the instability of the connection and the prototypical setup and configuration, turned out to be a characteristic problem for the E=MC^2 project:

- The line was effectively available for small time slots only, allowing no detailed observation, adaption of load balancing or application to increase the bandwidth exploitation, nor measurement repetitions.
- For days, all IP packets larger than 600 bytes oozed away between Paris and Berlin.
- Only 10Mbit/s access speed was available to the broadband lines.
- As described above, the latency was very large due to many high level protocol conversions.

- Only two workstations in Paris and just one in Stuttgart could be coupled fruitfully for this trial. This does not present real coupled workstation ‘clusters’. Hence, load balancing was encouraged artificially to distribute the computation over all hosts.

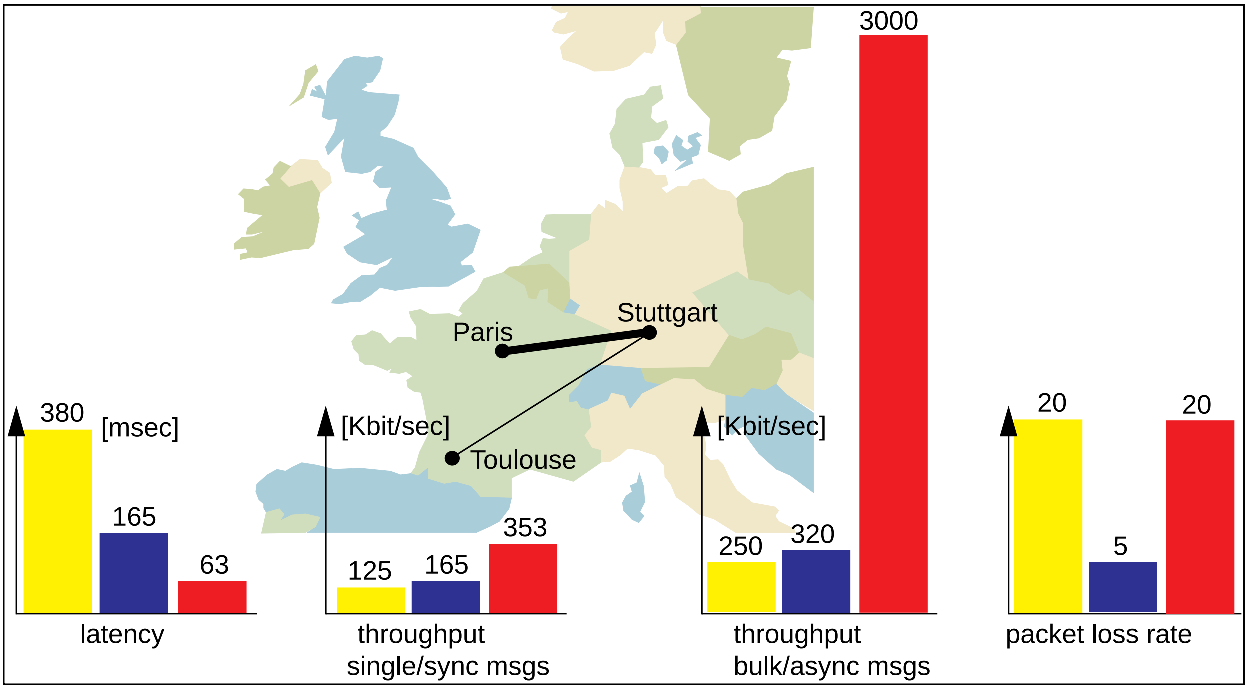


Figure 21: Network performance between Paris / Toulouse and Stuttgart.

One parallel finite element analysis application was performed. For comparison to the existing narrowband network, a similar configuration was set up and measured between Stuttgart and Toulouse, as shown in Figure 21. It should be mentioned, that the single, dedicated application, although parallelized, could not exploit the full bandwidth, but caused intermittently network traffic.

The execution profiles of the broadband configuration show acceptable ratios of data communication portion to CPU processing portion by full exploitation of all three machines with rather fine grained distribution of tasks. The workstation in Stuttgart initially owned all data, so large amounts of data had to be replicated and migrated across Europe during the element calculation. During equation solving it turned out that in-house data communication between the Interop machines was significantly less expensive due to smaller latencies. Load balancing was forced to distribute the tasks nevertheless. The execution profiles of the low speed network measurements show large communication overhead due to increased latencies and poor bandwidth.

Figure 22 compares the elapsed execution times for the element calculation phase and the average elapsed times per equation solving iteration:

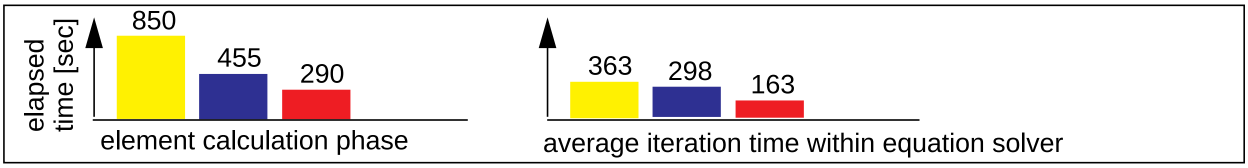


Figure 22: Application level performance of first European broadband trials.

Note that the evaluation of the broadband trials focussed on the load balancing and application level only. Network level performance evaluations to examine the effectively exploited bandwidth could not be performed due to lack of time.

The real obtained performance gain on the application level even under the adverse circumstances, together with the acceptable data communication overhead, show the following: The increased bandwidth can be reflected in corresponding speedup for distributed parallel applications; High speed networks can enable sufficient fine grained load balancing between European computing centers, provided advanced load balancing support and suitably decomposed applications, while current networks are insufficient.

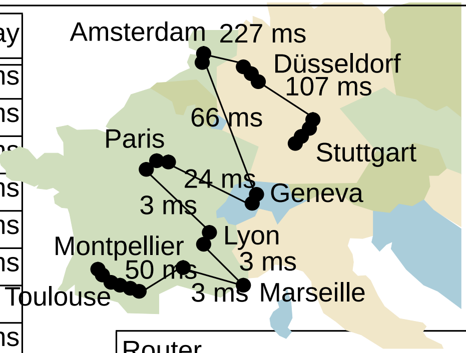
For larger application scenarios and multi user concurrence load balancing will be able to better exploit the network bandwidth. More workstations per cluster and more computing centers participated in most other measurements in order to enable real exploitation of European computing resources and match the trade-off between communication cost and utilization of CPU cycles, memory, disks etc.

6.5 Trans European Broadband Trials: Stuttgart - Toulouse

After an unexpected long time of organizational and technical problems, the ATM link between Stuttgart and Toulouse became available at least for several days. Still then, the reliability was not good.

The figure shows the typical message routing and latencies on the existing trans European network, and on the trans European broadband pilot connection established for the E=MC^2 trials between IPVR Stuttgart and CERFACS Toulouse.

Router	delay
gwinf.informatik.uni-stuttgart.de	4 ms
cisco1-ivaih.rus.uni-stuttgart.de	4 ms
cisco3.rus.uni-stuttgart.de	4 ms
Stuttgart1.BelWue.DE	5 ms
Duesseldorf3.WiN-IP.DFN.DE	52 ms
ipgate2.win-ip.dfn.de	107 ms
duesseldorf2.empb.net	
amsterdam7.empb.net	254 ms
Amsterdam1.dante.net	334 ms
Geneva1.dante.net	400 ms
Cern-EBS1.Ebone.net	
Paris-EBS1.Ebone.NET	
Renater-RBS1.Ebone.net	
stamand3.renater.ft.net	424 ms
stamand1.renater.ft.net	
lyon1.renater.ft.net	
marseille.renater.ft.net	
montpellier.renater.ft.net	
toulouse.renater.ft.net	
193.55.244.241	430 ms
192.70.80.206	
cerfacs-toulouse.octares.ft.net	
193.50.120.2	480 ms
mercure.cerfacs.fr	
frene.cerfacs.fr	510 ms



Router	delay
gwinf.informatik.uni-stuttgart.de	3 ms
cisco1-ivaih.rus.uni-stuttgart.de	4 ms
cisco3.rus.uni-stuttgart.de	4 ms
Stuttgart1.BelWue.DE	4 ms
(Datex-M routing invisible on IP level)	
153.17.200.248	43 ms
(ATM routing invisible on IP level)	
192.168.31.2	43 ms
mercure.cerfacs.fr	44 ms
frene.cerfacs.fr	44 ms

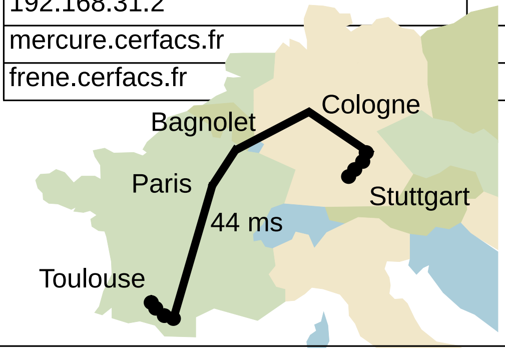


Figure 23: Current Internet vs. broadband pilot routing Stuttgart - Toulouse.

The performance comparison between the existing network and the trans European broadband pilot connection established for the E=MC^2 trials between IPVR Stuttgart and CERFACS Toulouse (Figure 24).

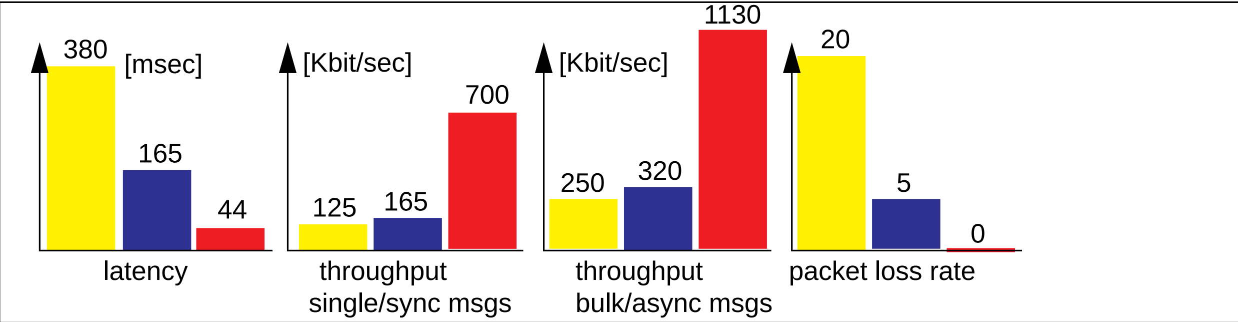


Figure 24: Trans European network performance between France and Stuttgart.

Using this broadband connection, detailed measurements involving all three application types could be performed and results were related to corresponding executions based on the narrowband connection. The figure depicts the network and host topology.

First, the three applications were launched in dedicated mode on the cluster. It turned out that, because all applications were originated at the IPVR, load balancing kept them almost within this one cluster. The reason for that decision was that the whole system was not too much overstressed, the evolving parallelism not too high, and the latency differences between local data communication and trans European communication were still large (1 msec vs. 44 msec). Hence, for a real exploitation of the meta computer for one single application, larger input data sets i.e. larger problems, would have to be calculated and the granularity of task interaction would have to be more coarse grained. Note, that the trans European trials between Paris and Stuttgart (described above) also investigated a single application scenario, but used a load balancing policy that was encouraged to distribute tasks more effusive.

Second, a multi user concurrence of 5 parallel finite element analysis applications was initiated at the IPVR cluster and executed under automatic load balancing support. The same scenarios where performed using the parallel image processing or the parallel database processing applications respectively. Figure 25 summarizes the application level performance for each of the application types, compared to the same trials running on the existing narrowband links at night or during a workday. These results built the most comprehensive evaluation of the question, how far automatic dynamic load balancing can exploit such European meta computer for typical HPC applications.

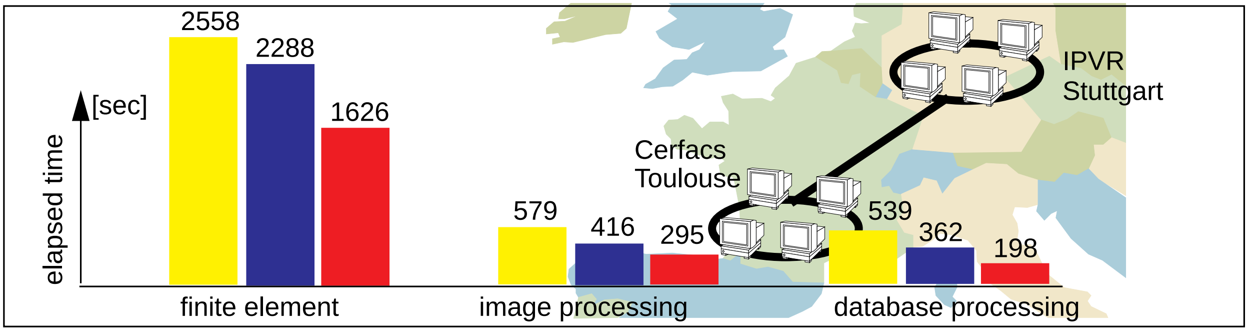


Figure 25: Application level performance and trial topology of second European broadband trials.

A more detailed look into the execution traces of the trials gives some interesting insights:

The parallel finite element analysis was distributed over the whole system on the broadband network and on the narrowband network as well. Figure 26 shows, that the communication portion of the tasks is considerably bigger on the slow network. Additionally, load balancing kept more whole applications on one cluster in the case of a poor network, sacrificing full compute resource exploitation in favor of data affinity, i.e. reducing expensive data communications. In the figure, each line represents the execution trace of one server, the time advancing from left to right. Black parts are task computations, while the grey part of the boxes gives the percentage of communication time spent. White areas are idle times.

The parallel image recognition applications (Figure 27) execute more fine grained and heavily communicating tasks in the first phases, and large sets of probably accessed common data in the further phases. Hence, load balancing did not dare to distribute applications across the WAN in the low bandwidth configuration. Moreover, with existing Internet it did not use the Toulouse cluster at all, because all applications originated at the IPVR, and at the time, when the IPVR cluster became overloaded by the applications, exploiting their parallelism, they already had created large sets of common data. So, the load balancing configuration that decided on per task base, mostly did not think it worthwhile to migrate tasks across Europe. In the broadband configuration, a better utilization of the meta computer was achieved. Still, only few applications were striped across nations and only for short phases. Mostly, load balancing tried to keep the applications within a cluster each, as far as a good overall system utilization was achievable.

Parallel database processing (Figure 28) again has rather different task granularity, parallelism and communication profiles. Load balancing exploited the distributed system in both cases. However, in the case of a poor network, the tasks within an application were kept more within one cluster each. This, similarly to the finite element application type, resulted in an overall worse resource utilization, and nevertheless significantly increased data communication portions for access to common global data. Another application type specific observation can be made here: Because the successor tasks within a query execution graph operate to a considerable degree on the intermediate data produced by their predecessor tasks, the migration of one task across cluster boundaries pulls the execution of related and succeeding tasks also towards the other cluster. In the figure, the low bandwidth network trace shows a shift from the IPVR where the applications started, to the CERFACS cluster. This is an execution at night; During working days, load balancing separated the applications into clusters.

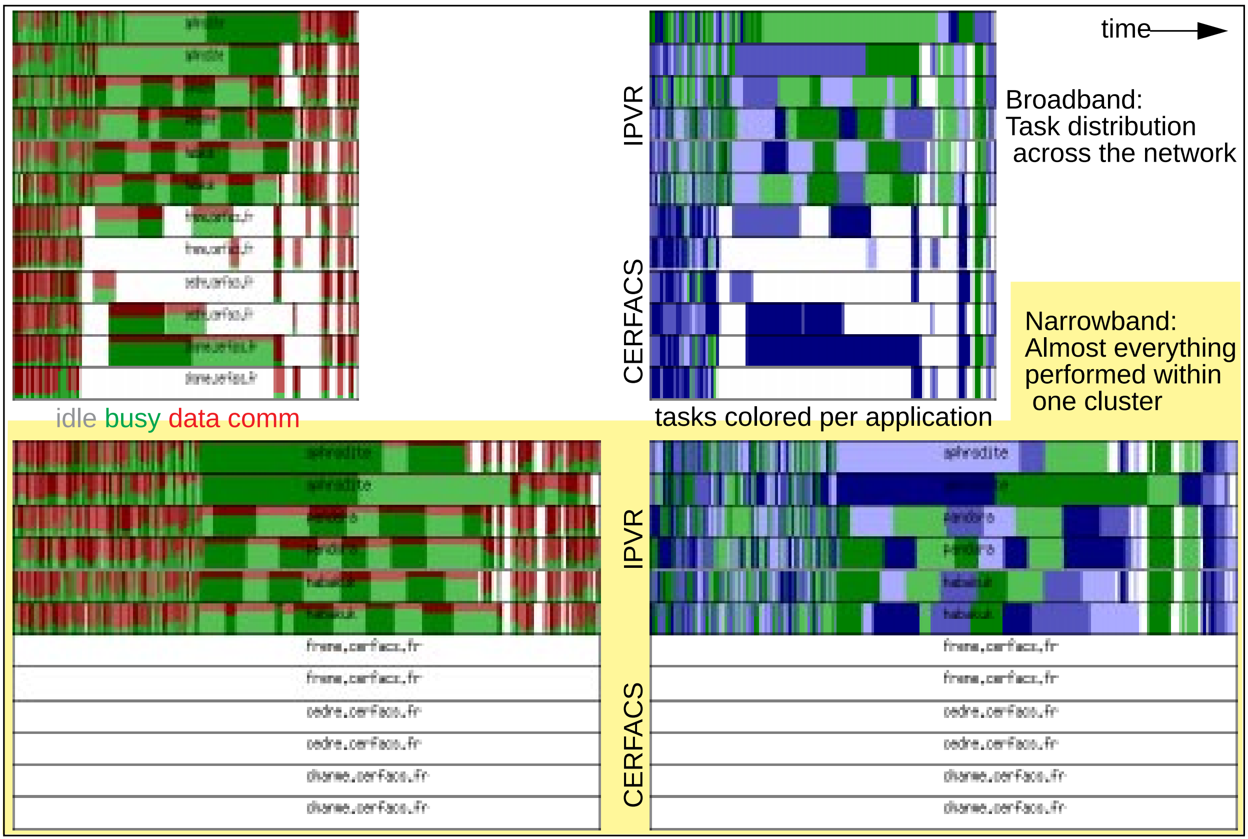


Figure 26: Execution profiles of concurrent parallel finite element calculations.

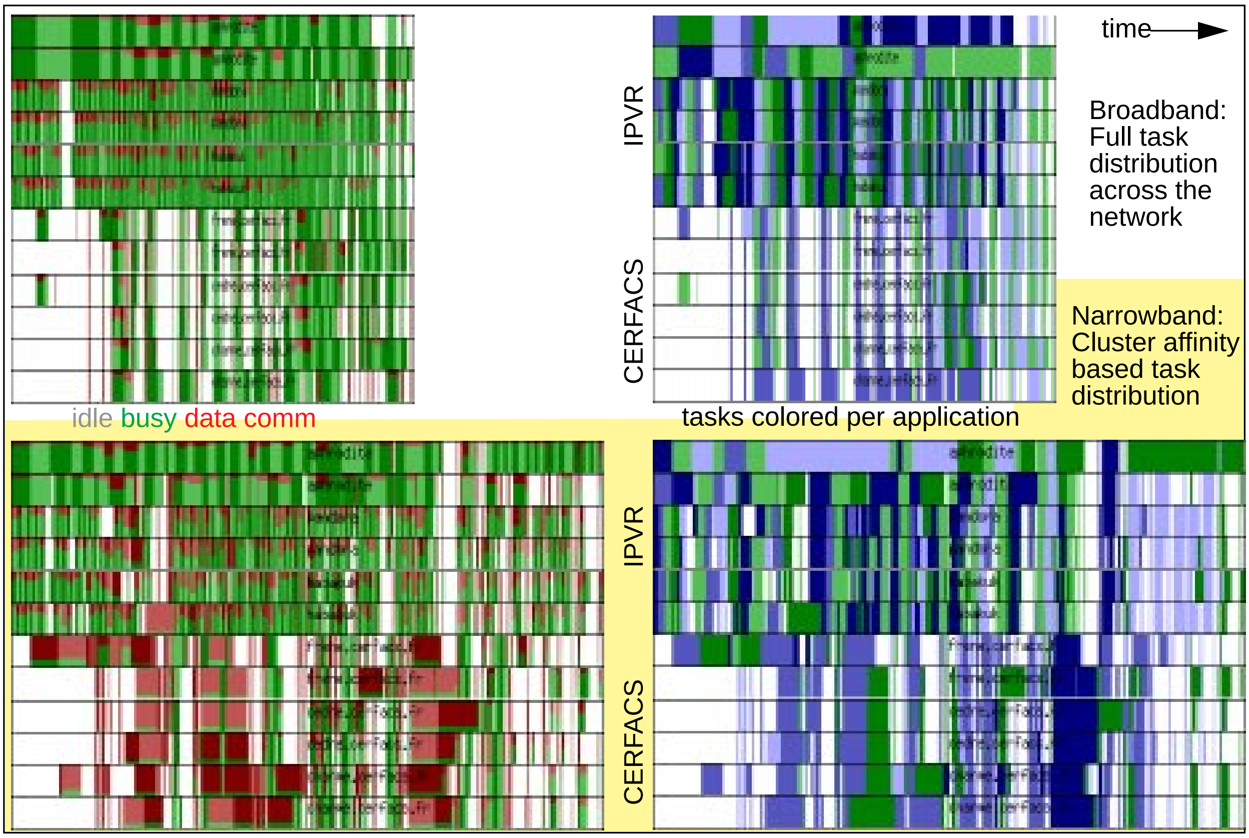


Figure 27: Execution profiles of concurrent parallel image recognition applications.

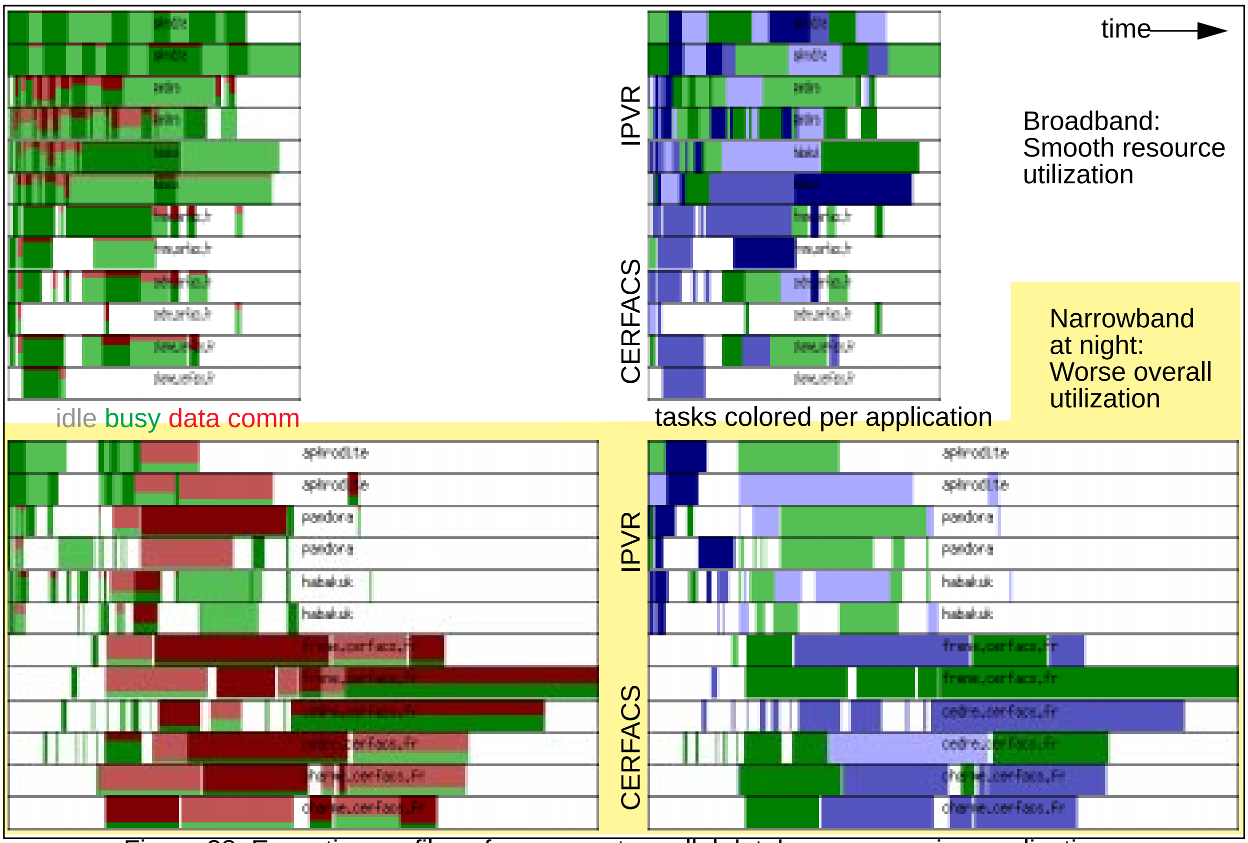


Figure 28: Execution profiles of concurrent parallel database processing applications.

6.6 Trans European Broadband Trials: Belfast - Stuttgart - Toulouse

Part of the network routing and performance within the European triangle was shown already in the previous sections. So, Figure 30 just depicts the typical message routing and latencies on the existing trans European network and on the trans European broadband pilot connection established for the E=MC^2 trials between IPVR Stuttgart and Queens Belfast.

The resulting network performance on this line is accordingly:

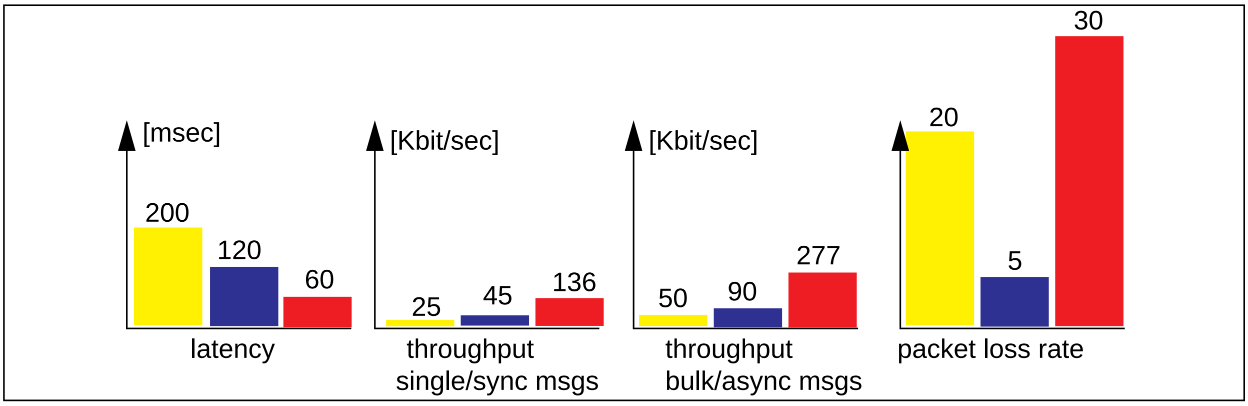
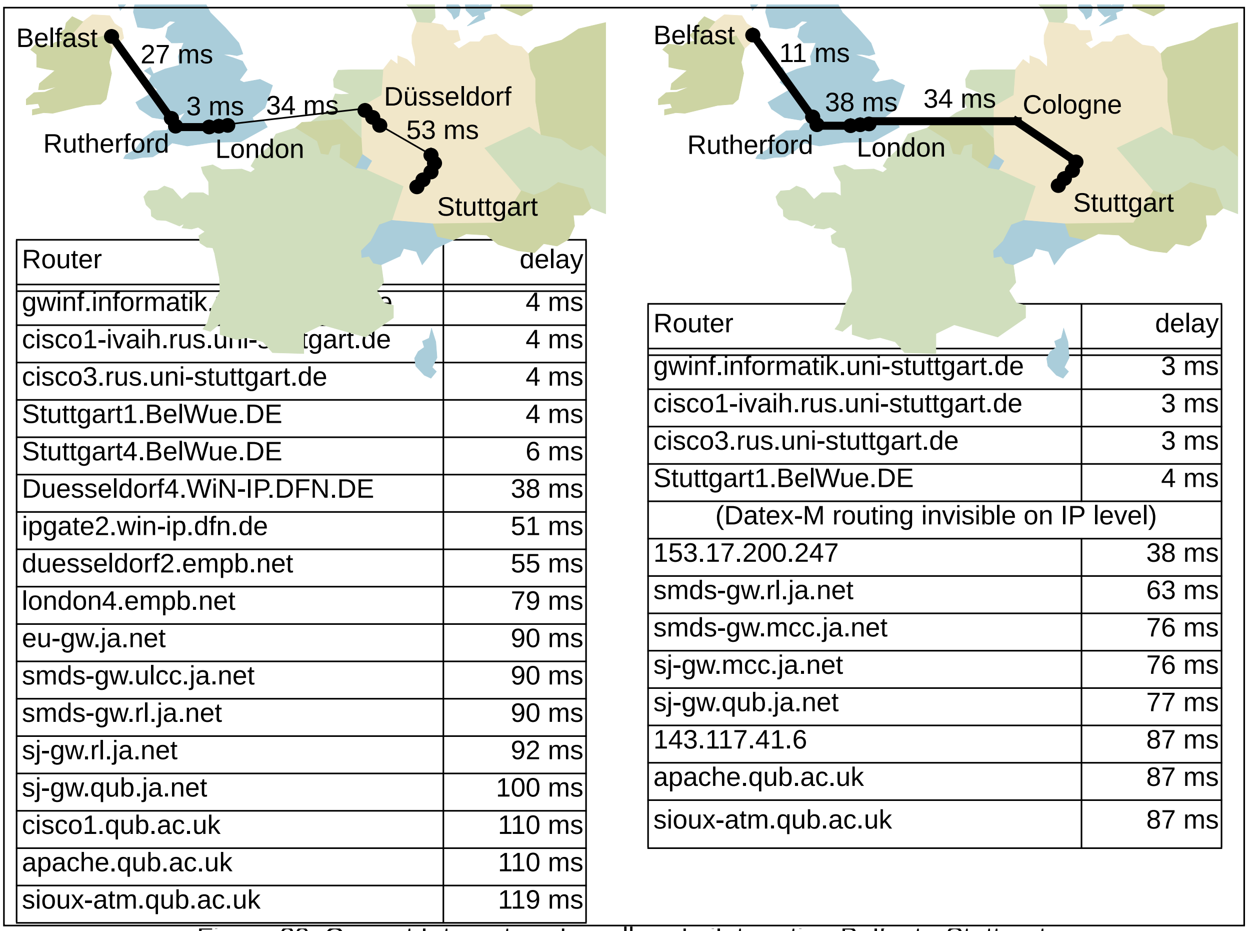


Figure 29: Trans European network performance between Belfast and Stuttgart.

The significant packet loss rate of 30% intimates misconfiguration of some switches along the broadband pilot connection.



Router	delay
gwinf.informatik.uni-stuttgart.de	4 ms
cisco1-ivaih.rus.uni-stuttgart.de	4 ms
cisco3.rus.uni-stuttgart.de	4 ms
Stuttgart1.BelWue.DE	4 ms
Stuttgart4.BelWue.DE	6 ms
Duesseldorf4.WiN-IP.DFN.DE	38 ms
ipgate2.win-ip.dfn.de	51 ms
duesseldorf2.empb.net	55 ms
london4.empb.net	79 ms
eu-gw.ja.net	90 ms
smds-gw.ulcc.ja.net	90 ms
smds-gw.rl.ja.net	90 ms
sj-gw.rl.ja.net	92 ms
sj-gw.qub.ja.net	100 ms
cisco1.qub.ac.uk	110 ms
apache.qub.ac.uk	110 ms
sioux-atm.qub.ac.uk	119 ms

Router	delay
gwinf.informatik.uni-stuttgart.de	3 ms
cisco1-ivaih.rus.uni-stuttgart.de	3 ms
cisco3.rus.uni-stuttgart.de	3 ms
Stuttgart1.BelWue.DE	4 ms
(Datex-M routing invisible on IP level)	
153.17.200.247	38 ms
smds-gw.rl.ja.net	63 ms
smds-gw.mcc.ja.net	76 ms
sj-gw.mcc.ja.net	76 ms
sj-gw.qub.ja.net	77 ms
143.117.41.6	87 ms
apache.qub.ac.uk	87 ms
sioux-atm.qub.ac.uk	87 ms

Figure 30: Current Internet vs. broadband pilot routing Belfast - Stuttgart.

Due to the late and unreliable availability of the pilot connections, only one broadband trial between Belfast, Stuttgart and Toulouse could be performed within E=MC^2 (Figure 31).

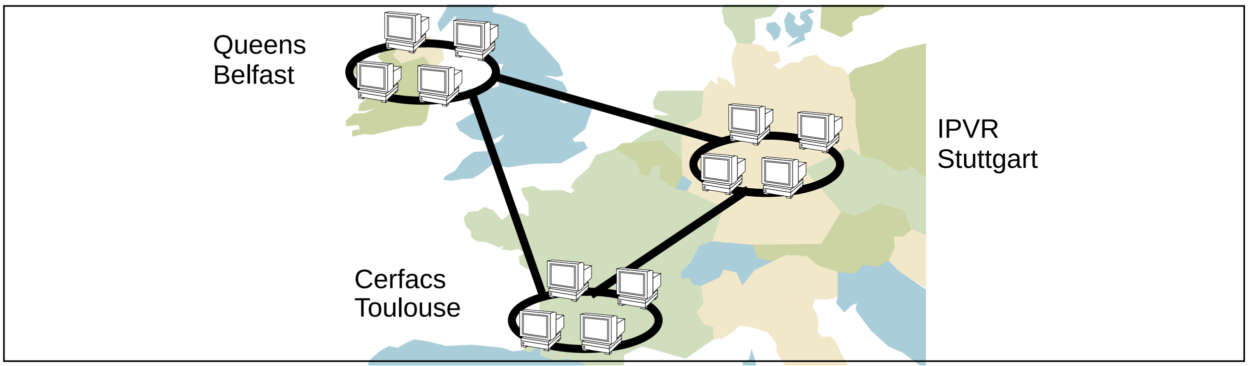


Figure 31: Network topology for third European broadband trial.

For this configuration, no comparable execution on the existing narrowband network was achievable, because the big packet loss rates, the huge latencies and the poor and rapidly varying throughput always led to message queue congestions or time-outs within some participating application processes. So, this measurement scenario mainly shows, that for heavy, real world load, and if actually several distant European clusters shall be coupled to form a meta computer, a certain quality of the network service is absolutely required. Figure 32 shows the broadband execution trace, mainly to illustrate the multitude of computational load within the distributed system. The lower part of the figure traces, to which application the tasks belong. This gives an impression of how much load balancing shifts tasks across the hosts and even clusters to achieve a continuous utilization of the resources. It must be stated that the load balancing policy which was employed here, often distributed more than suitable, i.e. introduced too much data communication cost that did not outweigh the exploited CPU time.

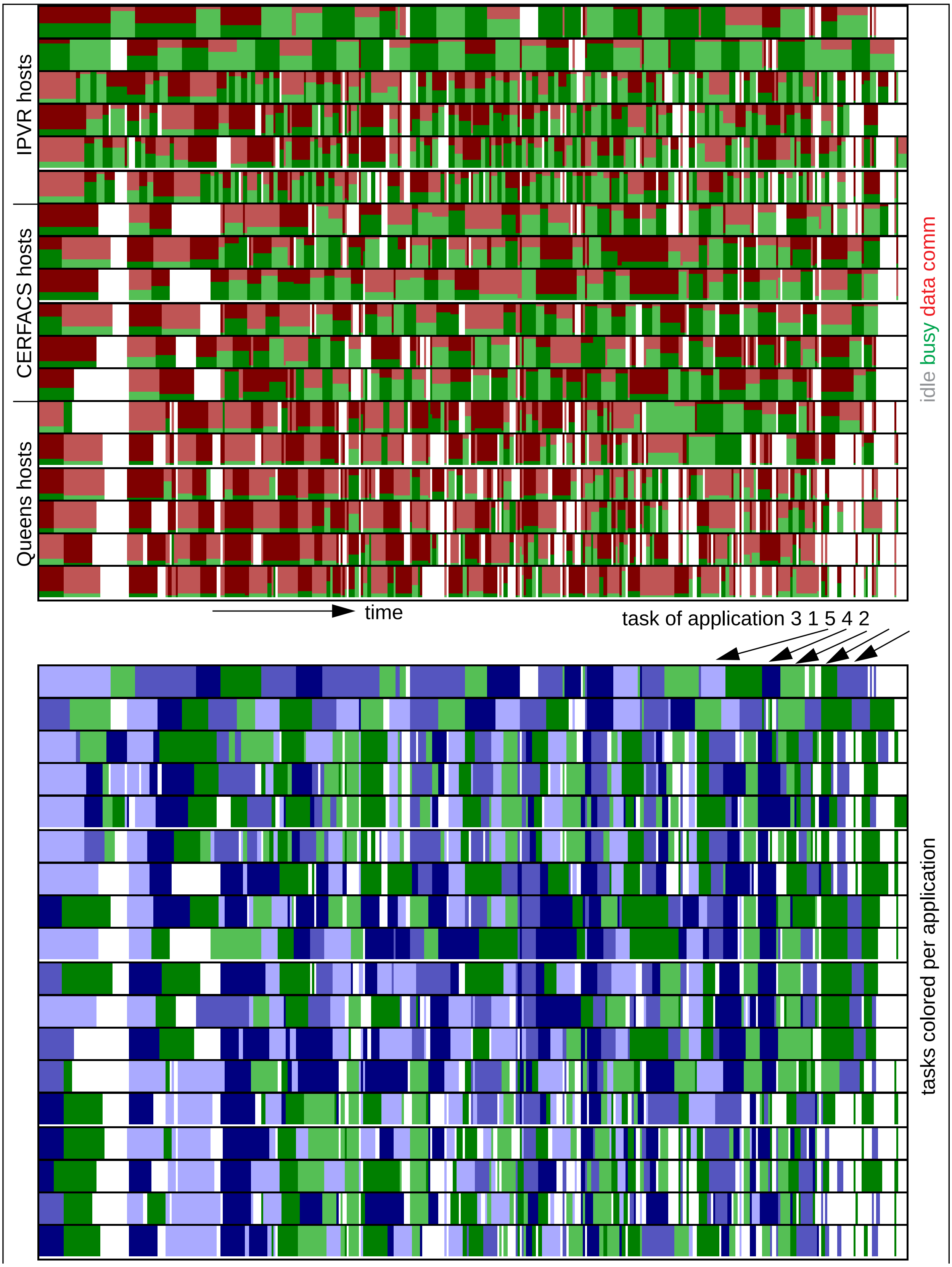


Figure 32: Execution profiles of third European broadband trial.

7 Feedback from Research Community and End Users

One important aspect of the project was the involvement of and feedback from end users and the common interest group. Hence, E=MC^2 attended international conferences, performed workshops and demonstrations and disseminated and discussed the approach and results by questionnaires.

7.1 Presentation at International Conferences

The E=MC^2 project presented and discussed its approach, besides presentations within earlier project phases, on the network fair ‘ATM developments’ in Rennes, France during March 1995 and on the HPC fair ‘High Performance Computing and Networking Europe’ in Milan, Italy during May 1995. In Milan, additionally a half day workshop presenting and evaluating the project was given, two papers concerning the project and the load balancing concepts were presented. Last not least, an on-line demonstration visualized five concurrent parallel image recognition applications were running distributed across Europe via ATM based broadband wide area network which was established between Milan and Stuttgart during the conference. Figure 33 shows a screen dump: The on-line visualization displays the current network traffic within and between clusters (large boxes), the host utilizations (small boxes) and the applications’ progress by printing a bar per finished task in the upper part of the window on one line per application.

The general opinion of most of the people to whom we talked in Rennes and Milan was, that the E=MC^2 meta computing approach and the actually performed trans European trials with complex relevant application codes, is very interesting. Some people even expressed, that the more well-known usage of broadband links, mainly for video transmission, is quite simple and well understood compared to efficiently distribute large HPC applications across coupled computing centers. Although, many people could not imagine that this way of distributed HPC is technically or even commercially feasible within the next few years. It still sounds like science fiction. Another interesting observation was, that some people where excited to see someone really demonstrating and investigating these issues, hoping that this could also help to change the Telecom pricing policies and push the public and political focus on these issues which were not doubted to become, not the most important, but another key factor of European technological competitiveness. Although some of the people were a bit disappointed that the E=MC^2 did not unravel dramatic new or unexpected discoveries or guidelines, all interlocutors agreed that it is worthwhile and interesting to strengthen and stabilize the possibilities, challenges, circumstances, requirements and limitations of trans European HPC by a variety of real important measurements, providing competent, quantitative evaluation results and experiences.

7.2 User & Interest Group Feedback to E=MC^2

Additionally to the questionnaire performed during the project definition phase, a questionnaire concerning the requirements towards remote execution facilities, and another questionnaire, discussing the project approach and providing early measurement results was broadcasted. It disseminated the E=MC^2 work in compact form to a variety of research and development people as well as application developers and end users around the world, and yield an interesting valuation and feedback from competent researchers and also from the user side of HPC and HPC centers. In the following, the aggregated questionnaire response from experienced researchers and developers in this computer science and scientific / commercial computing domain is presented. The character ‘+’ indicates that at least 80% of the responses confirmed the thesis, ‘?’ stands for mixed feelings and ‘-’ indicates prevailing disagreement.

- What investigations do users expect from E=MC^2?

  + technical feasibility of European wide area distributed HPC

  + bandwidth requirements / bandwidth utilization (parallel applications / multi user load)

  + potential benefits from trans European high speed networks

  + observation of application behavior / network utilization

  + user requirements and user satisfaction to rate marketability

- Relevance and importance of E=MC^2 project?

  + wide area distributed HPC is of important commercial benefit in the near future

  + for scientific, simulations

  + for database processing

  ? important compared to video conferencing, video transmission, cooperative work

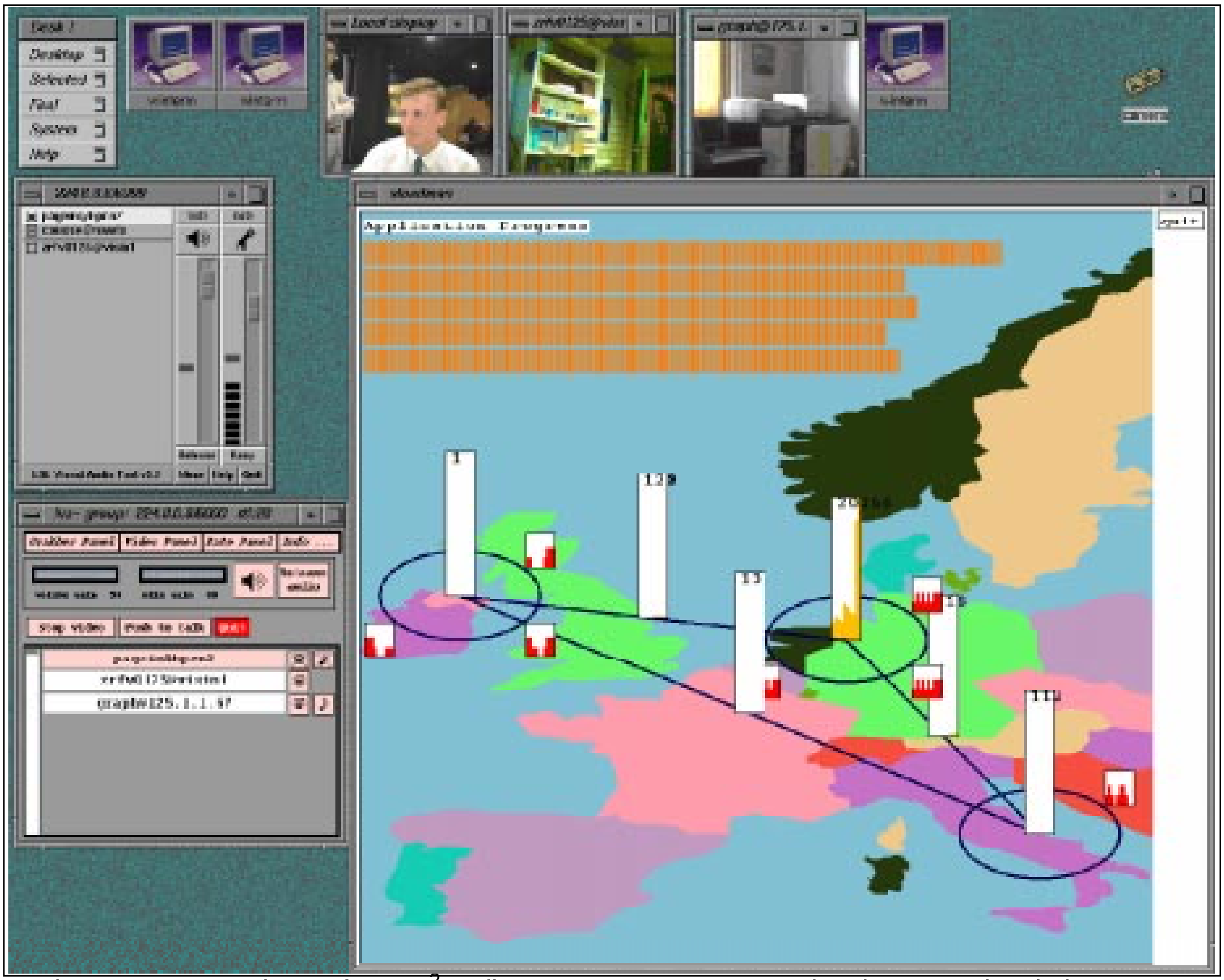


Figure 33: Screen dump of E=MC^2 on-line European Meta Computing demonstration during HPCN.

- User community that can profit from distributed HPC?

  + HPC centers

  + universities and research centers

  + companies that need huge computing resources

  - single end users for private applications

- Major challenges for distributed HPC?

  + network availability and reliability

  + network management

  + network bandwidth

  + network latency

  + network costs

  + operating system support for addressing/communication within applications

  + operating system support for automatic remote execution / load balancing

  + parallelization / restructuring / tuning of relevant applications

  + portability / flexibility of relevant applications

- Are the E=MC^2 trials suitable for the investigations?

  + heterogeneous computer architectures realistic

  + IP over ATM realistic

? IP over ATM only major way to use ATM for HPC next years

+ network and HPC topology realistic and interesting

+ application types relevant

? application types cover major HPC patterns

+ network monitoring relevant

- Load balancing trials - load balancing concept suitable?

+ dynamic task assignment, central & local task queues

+ no preemptive migration due to node heterogeneity

? centralized per cluster, decentralized between clusters

+ decision cost model: Minimum task response time (queue wait time + compute time + comm. time)

+ observe nodes’ load, data distribution, data communication cost

+ exploit pre-estimations of task sizes, data access patterns

+ load balancing explicitly considers communications

+ application task parallelized, client - server structured

- distributing fine granular communicating tasks can be profitable

? communication via virtually shared data is a suitable concept

- Load balancing trials - applications relevant?

Parallel finite element analysis:	+ realistic	- typical
Parallel image recognition:	+ realistic	? typical
Parallel complex database operations:	+ realistic	? typical

- Value of early $E=MC^2$ measurement results?
  - Characterization of existing networks

+ corresponds to common experiences

+ throughput and latency are the main characteristics

  - Comparison of current network night / day

+ scenarios suitable

+ results plausible

  - Trans European broadband trials between IPVR Stuttgart and Interop Paris

+ scenario suitable

+ results plausible

+ conclusions derivable

- Agreement to $E=MC^2$ general experiences

+ Large, important existing applications are inflexible, not ready for network distribution. Suitable application porting and tuning is necessary and expensive.

+ Internetworking configuration and addressing still difficult, wide area ATM even more complicated

? High speed networks are new technology and availability / stability are prototype like, no commercial strength

+ Operating system support has to be enhanced in terms of more transparent network management / addressing and automatic load balancing support

? On current low bandwidth networks across within Europe interactive remote work and coupled computing are infeasible

+ Distributing coarse grain large, isolated jobs already possible

+ Wide area coupled HPC is severely affected by concurrent multi user network traffic

+ Trans European high speed networks can be utilized profitable, they enable trans European cooperative computations and suitable workload balancing, i.e. the exploitation of the distributed computing resources like one large meta-computer.

+ Message latency is an important limiting factor for parallel and distributed HPC, if fine granular data communication and synchronisation within parallel applications is inherent to the problem.

+ A major part of important algorithms are bound to these closely coupled, communication intensive execution patterns

+ European meta computing is technically feasible

? European meta computing is commercially feasible with the approach investigated by E=MC^2

Some additional interesting comments are listed below:

- “E=MC^2 evaluation criteria are mainly performance issues, but there are also price and usability issues. The project work seems to focus on scientific computing which is less than 1% of the use of computers. I think that other applications (e.g. database, transaction processing, mail, interactive video) will be much more important.”
- Explicit agreement was expressed to the motivation and importance of the E=MC^2 project: “Investigation of another, economically very important and promising domain of applications using European high speed networks with completely different profiles and challenges than the more common and more understood domain of video/voice transmission and cooperative work.”
- Are network costs a major challenge? “Network costs are small, PTT prices and policies are HUGE.”
- Main expected user community: “HPC centers, universities and research centers are less than 5% of the society. This is less than sales tax in California. This sales tax is ignored and we can ignore these users too. The real users are companies and individuals. We would not build a special phone system for universities and HPC centers. We will not build a special net for them.”
- Americans view network latency and throughput of the existing European network as astonishing poor.
- Load balancing approach: “Load balancing should come from the application.”
- Choice of applications: “Heterogeneous applications are missing - Simulation and visualization.”
- “The commercial feasibility and profitability are important questions to be answered by projects like E=MC^2. If I knew the answers, the projects would not be necessary any more.”
- Communication via access to virtually shared data for database processing: “If the database application using virtually shared data is unchanged from its non-distributed version then it will perform poorly. However, a database designed explicitly to take advantage of virtually shared data, one that knows the cost, latency, etc, could perform well, possibly better than one based on a higher level client-server task decomposition model.”
- Major challenges for distributed HPC: “The network hardware will soon be available with sufficient quality, only the prices may cause problems. Most important is a ingenious management concept for proper ressource utilization.”
- Wide area coupled HPC affected by concurrent multi user traffic - solvable by ATM bandwidth reservation facilities? “Not quite sure: Fixed reservations on such networks that are used by so many endpoints concurrently, could drastically reduce the available bandwidth.”

8 Conclusion

We quantified the feasibility and limitations of current European networks to figure out the real need for trans European broadband networks. Therefore, the E=MC^2 project prepared and performed the trials as follows:

- The project partners identified appropriate, relevant existing applications, they did not start new development or specific porting and parallelization of software packages. However, the chosen applications had to be adapted for wide area distribution. They had to be carefully enhanced and flexibilized to run suitably on heterogenous processing nodes and across machines connected by heterogeneous networks.

- For all broadband trials, the IP over ATM protocol was employed rather than native ATM protocols, because all existing parallel / distributed applications are based on proprietary or on IP mechanisms.
- The participating high performance computing centers, PPC at Queens University Belfast, RUS and IPVR at University of Stuttgart, CERFACS at Toulouse, GMD at Bonn and RAL at Rutherford are currently connected by low speed networks (Ethernet & FDDI & ATM LANs and X.25 & leased lines for WAN). One main effort of the preparation phase in the E=MC^2 project was to achieve a pilot broadband connection, mainly based on the promising new ATM protocol to perform real international broadband trials across Europe.
- Compared to most other recent ATM broadband projects, the E=MC^2 project did not use ATM to connect two machines within a room to transfer continues data streams between them, but performed real complex applications distributed across the European wide area broadband network, namely between Bonn - Stuttgart, Paris - Stuttgart, Belfast - Stuttgart - Toulouse and Milan - Stuttgart. Hence, realizing European meta computing the project had to face various new challenges.

Despite the various technical problems and time limitations, the project managed to get some early, encouraging results. The presented measurements can be viewed as a rather pessimistic analysis due to the adverse circumstances. That means, less prototypical conditions will provide even more satisfying and promising results.

Several general observations and conclusions emerged from the project work:

- It is nearly impossible to work interactively on distant computing centers during day time. However, this is absolutely necessary to install software, to set up configurations and batch jobs as well as to watch, manage and control remote executions. At night or weekends it is possible although slow, but this is not enough for real scientific and commercial use.

  Wide area distributed computing clusters connected by existing low speed networks cannot be viewed as a large meta computer. There is a huge gap in terms of cooperation, communication and task exchange for balanced resource utilization between within a computing center and across centers.

  On current low bandwidth WANs it is not possible to perform load balancing on process, command or even on task level within big, parallel, mission critical applications; May be it is partially possible on a batch job level.
- The latency aspect was recognized as an important factor for parallel and distributed high performance computing on a fine grained level. At the first glance, latency seems to depend on the number of hops and conversions between the distant sites. However, network experts tell that the number of hops is no problem, neither for ATM nor IP: ATM switches consume about 0.01 msec to forward an ATM cell, IP router take about 1 msec per packet. In fact, the problem is caused by the congestion of the narrowband networks which results in long queues within the hops. The same problem is to be expected on the ATM networks once they are loaded similarly. Provided a suitable bandwidth reservation, latency will be finally dominated by the physical distance.
- The presented measurements can be viewed as a worst case analysis due to the adverse circumstances. For larger application scenarios and multiuser concurrence load balancing will be able to better exploit the network bandwidth. More workstations per cluster and more computing centers have to participate in further measurements in order to enable real exploitation of European computing resources and match the trade-off between communication cost and utilization of CPU cycles, memory, disks etc. Also different application types should be observed. Hence, further ATM measurements with direct high broadband access of the machines to the network are inevitable.
- On the application and load balancing level the broadband trials confirm, that parallel applications cannot be distributed arbitrarily fine grained, especially algorithms showing intensive data communication and synchronization. Broad band connected distant computing centers actually can be coupled to effectively increase the processing power, but it is important to decompose applications into coarse grained tasks as decoupled as possible. Latency bound communicating tasks should still be kept within local area clusters, data bound computations still close to their data. Hence, in principle load balancing can effectively distribute groups of related tasks within competing parallel applications across Europe, which can be parallelized in finer degree within the computing centers.

- Broadband connected computing resources throughout Europe can be fruitfully utilized by proper automatic, application independent load balancing support. Furthermore, the computing resources can be maximally exploited on rather fine grained task level. This enables to even dynamically spread large parallel applications across nation boundaries according to instant availability of network and computing capacities. However, the problems encountered with one of the most advanced and flexible load balancing environments showed, that much development work has to be done in this area to make suitable load balancing generally available. The difference between local area and wide area cooperation will not disappear.
- Preparing the complex, existing applications for distributed heterogenous processing showed that current parallel and distributed computing on workstation clusters suffers from insufficient concepts for naming, security, vendor independence and availability, which make remote processing, global file access, communication and management difficult:

  There is no useful or complete hierarchical naming scheme for countries, computing centers, hosts or network interfaces. Default directory paths for executables and users' homes are different and user names or identifiers often must be different. Transparent remote file access (e.g. NFS, AFS) is mostly not available across cluster boundaries, and if available then performance and reliability are poor. Illegal data access and resource usage across the network is prohibited by inflexible 'fire wall' routers at the computing centers that render remote execution and communication more difficult. Heterogeneous computer architectures and versions of the UNIX operating system differ in definition files and library functionality; They use different byte orders and byte alignments for data structure representation. Differing executables and data file formats must be managed additionally to the version management of programs, configurations and data across the network and complex scripts are necessary to manage and distribute the versions across the network. Each time, as network lines or participating hosts become unavailable, manual configuration changes are necessary.

  Because the UNIX operating system is not distributed, there is no transparent way for parallel computing and communication between distributed processes. Hence, due to poor operating system support for addressing, communication within applications, for automatic remote execution and load balancing, the parallelization, the restructuring and the tuning of relevant applications is quite difficult nowadays. Environments like PVM or DCE provide more transparency for distributed computing, but they became broadly available in the last two years and therefore still have performance and flexibility problems, remaining platform dependencies, and very few relevant applications obey these protocols so far.

  The difficulties in the many projects like EUROPORT that distribute and parallelize commercial applications in a presumably platform independent way, show that this is a subject of current development and that the existing applications are very hard to re-engineer. Even parallelized applications are fixed to run on a certain array of nodes within one homogeneous parallel machine only or within one local cluster. Applications are not easily portable and quite inflexible to be distributed arbitrarily. Finally, once an application is parallelized, it requires much time, effort and application as well as operating system experience to adjust task sizes, communication pattern and data layout, in order to make them run efficiently on a certain configuration.
- Specific problems that come along with distributed HPC when using wide area international broadband connections which can be summarized in following items:

  The comparably huge latency, the unreliability of both the broadband and the narrowband network and the relative low effectively visible bandwidth of the broadband lines as well as the current network cost made the trials less comprising and leave a gap between the trial results and the circumstances for real commercial usage.

  Distributed large HPC applications demand high bandwidth networks. They are not just isolated computations but large, parallelized, data communication intensive applications, showing heavily cooperating tasks and large portions of remote data access.

  A proper exploitation of the European computing resources requires not just keeping all processors busy, but also considering the imposed network load. Suitable task distribution granularity and clustering of tasks is necessary and must be as dynamically and automatically as possible.

Network Performance: Overall, the existing trans European network lines show very poor performance. The broadband links that have been set up during the project, show a large improvement in terms of latency and throughput. However, there are still many tuning parameters at the switches and routers which will have to be adjusted for optimum network performance. This can be seen sometimes at the packet loss rates due to suboptimal buffer sizing and message decomposition into frames and cells. In terms of throughput, the broadband links may reach conditions similar to usual LANs, but the latency will remain larger in orders of magnitude. The near future will provide more stable and efficient international connections, which should be able to reduce the long distance latencies by a factor of 10 compared to what is available today.

The broadband connection was intended to be available during 7 days continuously, but actually was available two of these days. This is characteristic for the state and strength of current broadband pilot connections beyond intra-campus networks. Not only to take commercial advantage but also to provide a useful service for research projects requires significantly more stability and availability. Another practical problem is the high latency as well as the huge gap between the raw throughput of the network and the sustained throughput on application level.

Currently, the main reasons for the poor availability are missing experience of the network providers concerning the new technologies, the difficulties of coordination between the network providers and administrators along the path and the poor hardware and software quality of the new network technology.

- For all trials the same team of hosts was used and the same number and sizes of applications were performed under the very same load balancing environment. The only differences are the characteristics of the underlying network which had to be learned by the load balancing facility itself by measurements and feedback from application behavior during runtime.

  Load balancing was able to squeeze remarkable throughput enhancements out of this coupled system by a sophisticated policy that exploited the network in a degree matching the trade-off between full CPU utilization and data communication costs. The one strategy successfully managed all three application types which show quite different profiles in terms of task granularity, potential parallelism, data communication frequency / message sizes and possible reference locality for global data access.

  Comparing the latency as well as the throughput differences between the narrowband and the broadband lines and the corresponding application throughput differences on these WAN coupled systems, it becomes obvious that, provided a suitable data, task and cooperation layout of the applications, and provided a suitable dynamic load balancing facility, the feasibility of building and using European meta computers strongly depends on the performance of the interconnection.

  On the application and load balancing level the broadband trials confirm, that parallel applications cannot be distributed arbitrarily fine grained, especially algorithms showing intensive data communication and synchronization. Broad band connected distant computing centers actually can be coupled to effectively increase the processing power, but it is important to decompose applications into coarse grained tasks as decoupled as possible. Latency bound communicating tasks should still be kept within local area clusters, data bound computations still close to their data. Hence, in principle load balancing can effectively distribute groups of related tasks within competing parallel applications across Europe, which can be parallelized in finer degree within the computing centers.

  Broadband connected computing resources throughout Europe can be fruitfully utilized by proper automatic, application independent load balancing support. Furthermore, the computing resources can be maximally exploited on rather fine grained task level. This enables to even dynamically spread large parallel applications across nation boundaries according to instant availability of network and computing capacities. However, the problems encountered with one of the most advanced and flexible load balancing environments showed that much development work has to be done in this area to make suitable load balancing generally available. The difference between local area and wide area cooperation will not disappear.

Some global observations from the load balancing trials corroborate common, intuitive expectations by these real measurements:

- Multi user HPC scenarios, i.e. computational load profiles with different unrelated, concurrent tasks / applications, are closer correlated to the network bandwidth, while the dedicated application executions depend more heavily on the network latency.
- In local networks with lower latency by more than an order of magnitude, network improvements do not result in according speedup of HPC applications, because the latency does not differ so much, and the latency is the major limiting factor there.
- The strong correlation of end user’s application performance to the underlying network power which still holds for really high bandwidth lines, gives rise to the usefulness and absolute necessity to establish trans European broadband lines for this domain of HPC. However, it must be kept in mind, that propagating the network enhancement up to the user requires serious, non-negligible efforts to adjust the applications and integrate load distribution support into operating systems.

The ongoing research within the E=MC^2 project concentrates on more detailed measurements, extended scenarios, more detailed network monitoring and closer end user integration. A prototype of a flexible brokerage service for end users and computing center administrators is the next step towards real exploitation of the European high performance computing resources for commercial and scientific purposes.

Acknowledgments

The measurements could be achieved with the help of many networking and computing engineers and scientists within and associated to the project. Especially, Paul Christ, Holger Fahner, Peter Feil (RUS Stuttgart), Lothar Klein, Petra Link, Roland Völpel (GMD Bonn) helped organizing and administrating these trials. The project team Kevin Brooks, Andrew Hilborne, David Horne, Paul Huish (Octacon Middlesborough), Rick Mc Connel, Richard Rankin, Paul Sage (Queens Belfast), Franz Fabian, Uwe Geuder, Walter Strommer (IPVR Stuttgart), Frederic Carbonnel, Patrick Laporte (CERFACS Toulouse), Ali Kabou, Fabrice Magnan, and Hubert Zeyssolff (Telmat Soultz) planned, organized and performed the trials, management and publications. Many researchers, scientists and engineers responded to the questionnaires and significantly contributed by interesting comments and discussions.

References

[Beck93] W. Becker, *Globale dynamische Lastbalancierung in datenintensiven Anwendungen*, Faculty Report 1993-1, University of Stuttgart, Institute for Parallel and Distributed High Performance Systems, 1993

[Beck94a] W. Becker, *Das HiCon-Modell: Dynamische Lastverteilung für datenintensive Anwendungen auf Rechnernetzen*, Faculty Report 1994-4, University of Stuttgart, Institute for Parallel and Distributed High Performance Systems, 1994

[Beck94b] W. Becker, R. Pollak, *Efficiency of Server Task Queueing for Dynamic Load Balancing*, Faculty Report 1994-9, University of Stuttgart, Institute for Parallel and Distributed High Performance Systems, 1994

[Beck94a] W. Becker, G. Waldmann, *Exploiting Inter Task Dependencies for Dynamic Load Balancing*, Proc. IEEE 3rd Int. Symp. on High-Performance Distributed Computing (HPDC), San Francisco, 1994

[Beck94b] W. Becker, J. Zedelmayr, *Scalability and Potential for Optimization in Dynamic Load Balancing - Centralized and Distributed Structures*, Mitteilungen GI, Parallele Algorithmen und Rechnerstrukturen, GI/ITG Workshop Potsdam, 1994

[Beck95a] W. Becker, *Lastverteilung in Workstation-Netzen*, BI Special Issue on Parallel Computing, RUS, University of Stuttgart, 1995

[Beck95b] W. Becker, *Das HiCon-Modell: Dynamische Lastverteilung für datenintensive Anwendungen auf Rechnernetzen*, Informatik Forschung und Entwicklung Vol. 10 No. 1, Springer Verlag, 1995

[Beck95c] W. Becker, G. Waldmann, *Adaption in Dynamic Load Balancing: Potential and Techniques*, Tagungsband 3. Fachtagung Arbeitsplatz-Rechensysteme (APS), 1995

[Beck95d] W. Becker, *Dynamic Balancing Complex Workload in Workstation Networks - Challenge, Concepts and Experiences*,Proc. Int. Conf. High Performance Computing and Networking (HPCN) Europe, LNCS, Springer Verlag, 1995

[Beck95e] W. Becker, *Dynamische adaptive Lastbalancierung für grosse, heterogen konkurrierende Anwendungen*, Ph.D. Thesis, University of Stuttgart, Institute for Parallel and Distributed High Performance Systems, 1995

[Blum94] R. Blumofe, D. Park, *Scheduling large-scale parallel computations on networks of workstations*, Proc. High Performance Distributed Computing, USA, 1994

[Buba94] M. Bubak, J. Moscinski, R. Slota, *Implementation of parallel lattice gas program on workstations under PVM*, Parallel Scientific Computing, LNCS 879, Springer Verlag 1994

[Cap93] C. Cap, V. Strumpen, *Efficient parallel computing in distributed workstation environments*, Parallel Computing, Vol. 19, No. 11, 1993

[Colb95] A. Colbrook, J. Elliott, M. Lemke, F. Wray, *EUROPORT2 - ESPRIT European Porting Action No 2*, HPCN Europe, Italy, 1995

[EMC94a] P. Huish (Ed.), *European Meta Computing Utilising Integrated Broadband Communications - Trials & Applications Descriptions*, Deliverable CEC Project B2010 TEN-IBC E=MC^2, 1994

[EMC94b] P. Huish (Ed.), *European Meta Computing Utilising Integrated Broadband Communications - Evaluation Plan*, Deliverable CEC Project B2010 TEN-IBC E=MC^2, 1994

[EMC95a] P. Huish (Ed.), *European Meta Computing Utilising Integrated Broadband Communications - Interim Report*, Deliverable CEC Project B2010 TEN-IBC E=MC^2, 1995

[EMC95b] P. Huish (Ed.), *European Meta Computing Utilising Integrated Broadband Communications - Evaluation Report*, Deliverable CEC Project B2010 TEN-IBC E=MC^2, 1995

[EMC95c] E=MC^2 consortium, *The European Meta Computing Utilizing Integrated Broadband Communications (E=MC^2) Project*, Proc. Int. Conf. High Performance Computing and Networking (HPCN) Europe, LNCS, Springer Verlag, 1995

[Gosc91] A. Goscinski, *Distributed Operating Systems: The Logical Design*, Addison-Wesley, 1991.

[Hand91] R. Handel, M. Huber, *Integrated Broadband Networks - An Introduction to ATM-Based Networks*, Addison-Wesley, England, 1991

[Horn94] D. Horne, *The E=MC^2 Project*, ULCC NEWS, University of London Computing Centre, 1994

[Lin94] M. Lin, J. Hsieh, D. Du, J. Thomas, J. MacDonald, *Distributed Network Computing over Local ATM Networks*, Proc. Supercomputing ‘94, USA, 1994

[McCa94] J. McCabe, *ATM in a Supercomputer Network Environment*, Tutorial, Supercomputing ‘94, USA, 1994

[Mech95] C. Mechoso, *High-Performance Computing and Networking for Climate Research*, HPCN Europe, Italy, 1995

[Mier95] H. Mierendorff, K. Stüben, C. Thole, O. Thomas, *Europort-1: Porting industrial codes to parallel architectures*, HPCN Europe, Italy, 1995

[Pozz95] E. Pozzetti, G. Serazzi, *Characterizing the resource demands of TCP/IP*, HPCN Europe, Milan, 1995

[Rahm93] E. Rahm, R. Marek, *Analysis of Dynamic Load Balancing Strategies for Parallel Shared Nothing Database Systems*, Proc. 19th VLDB Conference, Ireland, 1993

[Roy94] M. Roy, *TEN-IBC ‘95*, European Commission Directorate XIII B Advanced Communications Technologies and Services, RA947242, 1994

[Stru95] V. Strumpen, T. Casavant, *Exploiting communication latency hiding for parallel network computing*, Int. Conf. on Parallel and Distributed Systems, Taiwan, 1995

[Tane85] A. Tanenbaum, R. van Renesse, *Distributed Operating Systems*, Computing Surveys, Vol. 17, No. 4, 1985

[TEN94a] TEN-IBC, *TEN-IBC Concertation Proceedings*, European Commission Directorate XIII B Advanced Communications Technologies and Services, 1994

[TEN94b] TEN-IBC, *TEN-IBC Baseline Document*, European Commission Directorate XIII B Advanced Communications Technologies and Services, RA947538, 1994

[Tolm95] D. Tolmie, *Gigabit LAN issues - HIPPI, fibre channel, or ATM?*, HPCN Europe, Italy, 1995

[Wolm94] A. Wolman, G. Voelker, C. Thekkath, *Latency Analysis of TCP on an ATM Network*, Proc. WinterUSENIX Conference, USA, 1994