

Xibalba PC Cluster Infrastructure

Department of Computer Science

Prof. H.J.Schek (Power Databases)

Prof. G. Alonso (Workflows)

Prof. K. Nagel (Traffic Simulation)

Prof. T. Stricker (Systems Architect)

<http://www.xibalba.ethz.ch/>



Cluster of PCs - A computational platform for computer science research

Clusters with a large number of low cost personal computers are gradually replacing traditional vector supercomputers. Clusters of PCs benefit from the market volume of the consumer market to achieve high compute power at an extremely low cost. During the past five years several groups in the department of computer science have related some of their research to this new platform by working on the technologies and the design of such systems, by adapting their work in databases to clusters, by investigating the scheduling of tasks and workflows in scientific computation on clusters or simply bringing, the important application of car traffic simulation to clusters of PCs.

Building a common, shared infrastructure

As it became clear that several groups needed a cluster for their research the issues of the minimal size and the required architectural characteristics were raised and discussed with the inhouse cluster architects. Despite the fact that the communications requirement for all application codes involved was within a narrow range, each group aimed for the largest cluster they could afford to experimentally prove the scalability of their ideas to large systems. It became apparent that a clever sharing concept for this

research infrastructure would result in access to a much larger system and open a unique opportunity.

Mode of operation in computer science

Many uses of computers in computational science just ask for readily and cheaply available compute cycles, that can provided by any infrastructure, regardless whether it is operated by the research group itself, by a university computing facility or by a national supercomputer center. The requirements of computer science researchers are



Figure 1: Front view of the Xibalba PC Cluster, the 128 nodes are mounted in 9 cabinets.

quite different. In many computer science research projects the compute platform itself, including its hardware and software, is part of the experiment and needs to be controlled by the researchers. Such a mode of operation is largely incompatible with the setup of supercomputer center that provides access on a “per job” and not on a “per machine” basis.

Furthermore the planned research in power databases requires a powerful I/O

system in each cluster node, which is usually not available in clusters designed for scientific computing. With the Xibalba cluster concept all four participating research groups could bring their hardware and software requirements into the project. The resulting system remains flexible after its installation and the groups are welcome to contribute new, additional system software with their experiments.

Dealing with different operating systems - multi-boot means multi-purpose

The different groups working on the Xibalba cluster do rely on vastly different operating systems and different middleware packages. The power database group works with Windows 2000 and the SQL Server database management system provided by Microsoft Corporation in a research agreement. The other groups do work with different flavors of LINUX, which can be quite specifically configured to their experiments. The architecture group maintains its minimal service operating system for diagnostics and maintenance and a setup with experimental communication system software for benchmarking and application with particularly high communication demands (high bandwidth and low latency). To support all the different needs the Xibalba system is conceived as a multi-boot system. Each node can be configured to boot in any of the operating system configured. In addition to the pre-installed operating systems the cluster is equipped with “Dolly”, a specialized software distribution tool that uses Xibalba’s powerful networking infrastructure to distribute entire new software installation within minutes.

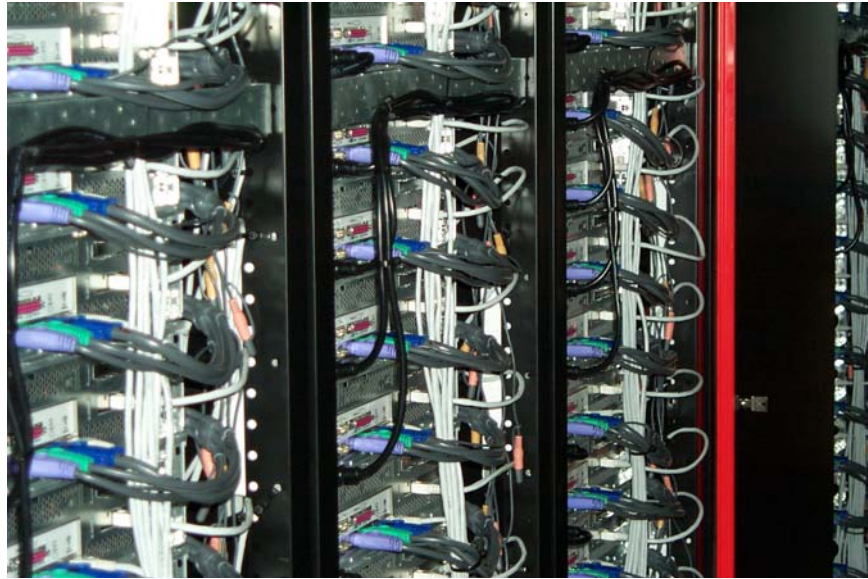


Figure 4: Rear view of the Xibalba PC Cluster - wires of 3 different networks connect the 128 nodes.

The technical data of the Xibalba cluster

The core of the Xibalba cluster are its 128 Pentium III based personal computers running at 1GHz. Some nodes do have dual processors, but the memory is kept uniformly at 256 MB per processor. High end motherboards provide a memory system with excellent characteristics. For the planned database work some special consideration was given to secondary storage in Xibalba. Each node includes two fast 7200 RPM ATA disk drives and two 10000 RPM disk SCSI drives with a capacity of 80 GByte to provide storage for operating systems, scratch space and for user databases at an optimal cost-performance ratio. The total storage of the cluster is over 10 Terabytes. In search of the optimal cost/performance tradeoff the inter-processor communication facility is the most critical part of a cluster. The networks of Xibalba are based on commodity 100 MBit/s Ethernet interconnects, like in most Beowulf class

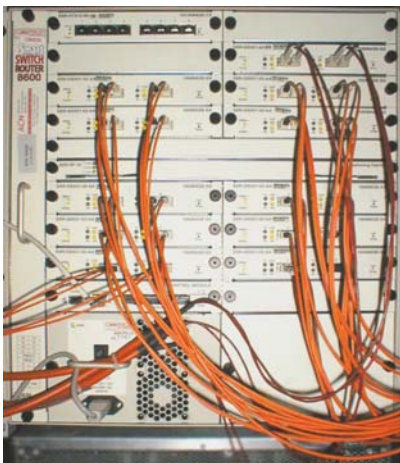


Figure 2: Cabletron SSR 8600, the type of fiber-optic Gigabit Ethernet router/switch as used in The CoPs cluster and the backbone.

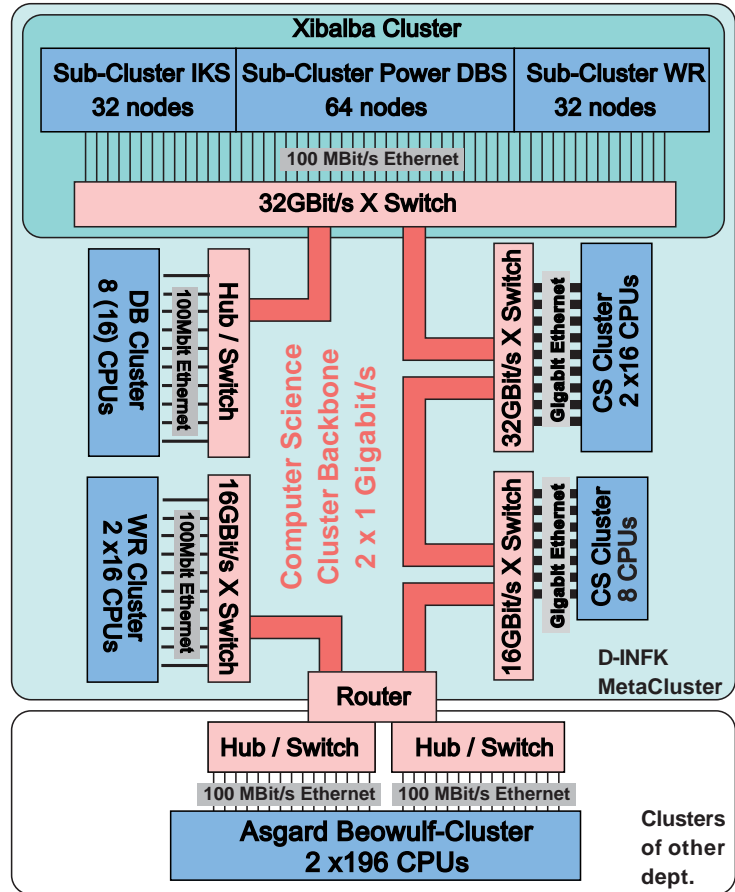


Figure 3: Topology of the Cluster backbone linking the different clusters of the department of computer science. A series of 2 x 1 Gigabit/s high speed interconnects link five departmental clusters with different performance characteristics and the Asgard Beowulf cluster of the university that is located in the same building as Xibalba. Among the different clusters the CoPs cluster of the parallel and distributed systems group provides an exceptionally strong communication system with a true Gigabit/s connectivity to every single node in the cluster. This cluster serves as researchplatform for new high speed interconnect technologies and as a testbed for applications with high demand on interprocessor communication. The cost of the equipment installed in this infrastructure is about one million Swiss Francs.

systems, but there are several small differences: For guaranteed compatibility with all current and future operating systems the keyboard, video and mouse signals are switched to a central console. A secondary low performance communication system provides guaranteed and reliable access to all nodes in all modes of operation and separates

the traffic of system and user communication. The primary data network is specified to sustain full speed non blocking communication on all ports simultaneously. With advanced system software the communication latency can be reduced to an absolute minimum. After the successful installation and initial testing the Xibalba cluster was benchmarked by its architects with the standard LINPACK benchmarks used to establish the top 500 list of the fastest computers in the world. A initial, unofficial measurement resulted in the performance of 55.5 GigaFlop/s Rmax which had put Xibalba at rank 497 into the published list of November 2000.

Computer Science Cluster Network Backbone

All cluster computing facilities in the department of computer science are linked together with a dedicated fiber-optic backbone. This networking infrastructure is in addition to the regular networking facilities provided by the university for general campus computing. Since the department operates several clusters with different characteristics it is important that the different experimental computations can access the central services of all groups involved. The communication traffic of distributed computations is burstier than normal network use and takes its infrastructure often to the limits. Therefore a separate cluster communication infrastructure is required to provide a predictable environment and to exclude interference with other uses. With this infrastructure all our clusters can be pooled together to form one big computational grid

Systems software engineering in Xibalba

The difference in equipment cost between a traditional supercomputer and a high performance clusters is reflected in the experience of engineering quality we gained during the construction of Xibalba. A noticeable imbalance between arithmetic units, memory system and communication system is to be compensated by system software, which remains a major challenge in the world of commodity PC components and free operating systems. Still - after a short period of juggling with different firmware, driver versions and network configurations the Xibalba cluster became a stable system that passed the all acceptance tests.