QTP Computing Laboratory Strategy

Erik Deumens
Quantum Theory Project
12 September 2001

Introduction

From the beginning of its computer operations (1980-1982) QTP has worked from a strategy and an architecture to realize that strategy. The motives were and are value, effectiveness, maintainability and security.

In the Spring of 1999, QTP updated its strategy for computing. This document is an update that addresses some of the issues that have arisen since then. The changes are refinements and clarifications. They do not reflect any fundamental change in vision. A reasonably comprehensive strategy for the next decade now can be formulated. The strategy has as the central theme the responses to the question of capacity versus capability in each of a few focus areas relevant to QTP computing.

Ten to twelve years ago, this question had an answer that in retrospect seems simple. Both capacity and capability were provided by a single hardware architecture. That strategy presented a uniform image to the user by providing all services in all focus areas via a single software and hardware solution. QTP used that strategy to the limit and achieved an efficiency of productivity versus dollars and effort that experience suggests is higher than commonplace in Computational Chemistry and Physics laboratories throughout the world. However, marketplace developments and QTP's success at optimizing the strategy have made further evolution along the same line impossibly difficult. A fresh strategy is required.

Current and future QTP system architecture

A computing strategy is embodied in architecture. The fresh strategy will consist of the architecture about to be discussed. Some parts are already in place, some are about to be put in place, some items are recommendations for action as soon as possible. The basic principles and reasoning by which this architecture is obtained are given in later sections. Major elements of the architecture are as follows.

- A firewall and gateway station and the use of private IP subnets 192.168.1-127.* to isolate internal space from external space and facilitate effective security and performance management. The private IP subnets are blocked by NERDC at the building Point-of-presence, as required by the IP standard. Thus hosts on private IP networks cannot be reached from outside the building. They can only reach outside of the building, through a proxy server. As a result, the QTP system acts to the outside world as a single object, which enables us to respond quickly and efficiently to requirements posed by UF (like authentication standards) and opportunities offered at UF (like Internet II). **Status** - Currently in place in the form of a SUN UltraSPARC workstation running Solaris

- A central server with multiple subnets directly attached to get optimal network performance to the desktop and with large reliable storage. Direct network connections rather than relying on external routing equipment reduces network overhead and makes internal operation of the system independent of an externl network. **Status** - The network currently in place is switched 10BT; to be upgraded during Winter 2002 to switched 100BT. The server is a SUN Enterprise 5000 running Solaris. The central storage is a 200 GB RAID system, to be extended durning Winter 2002.

- Desktop stations with large high-quality color display as their main feature and requirement, for scientists and administrative staff. Scientific and administrative staff have different primary software requirements and thus use of different OS is warranted and advised for stability and simplicity of support. **Status** - Since May 1999 desktops are a mixture of Pentium and UltraSPARC stations running Solaris for scientific staff and Pentium stations running NT 4.0 for all administrative staff.

- One or more Pentium servers running NT 4.0 Terminal Server to provide stable access to commodity software within the secure, flexible UNIX environment. **Status** - One dual processor P-II 260 MHz in place since March 1999. To be upgraded and maybe expanded the Fall of 2001.

- Cluster of production compute nodes providing CPU cycles for standard computations. Each node should have a fast CPU, reasonably sized RAM, and fast disks. The principle factor determining what hardware to buy for the cluster nodes (presuming operating system compatibility and compliance with proper security and manageability standards) is to maximize the capacity, i.e. CPU cycles, per dollar invested. **Status** - Xena phase I (134 node RS/6000 SP) came online in Nov 2000. Xena phase II (176 node RS/6000 SP) is expected to come online Nov 2001.

- One or more fast development nodes for interactive access in sessions of one to several hours, in addition to the central server, which should be used for development and debugging. **Status** - Quanta (16 node RS/6000 SP) in place since 1998. Quanta phase II (61 node RS/6000 SP) is expected to come online Nov 2001.

- One or more graphics stations to provide high quality visualization and video capability. **Status** - Together with Physics, a lab has been built in NPB 1213 with nine IBM RS/6000 42P model 170 and 270 stations for visulaization and simulation and an SGI Onyx 2000 with a FakeSpace Immersadesk R2 for 3D visualization.

- One special purpose compute node with very large RAM for memory intensive production jobs. **Status** - Four nodes in Quanta provide this capability.

- One parallel computing machine with mixed distributed and shared memory architecture consisting of several nodes that are themselves symmetric, shared memory multiprocessors connected by a low latency, large bandwidth communication switch. This machine is for development, validation, and tuning of super-computing software and for limited scale production with that software. It should be kept state-of-the-art

to allow QTP software again to be relevant in the super-computing arena. It can provide extra capacity of CPU cycles, but only at low priority. **Status** - Ten nodes of Quanta provide this capability.

- One Beowulf style compute cluster to experiment with the cost effectiveness of this modern technology. **Status** - Expected Fall 2001 or Winter 2002.

This architecture requires that QTP supports at least two operating systems: Solaris and some mature variant of Microsoft Windows. In addition, AIX, Linux and Irix are used on research computing equipment. To simplify security administration, all operating systems except Solaris will be on private IP subnets, thus making them invulnerable to oitside probes and attack.

- Solaris is a secure, stable, flexible, network-centric, high-performance operating system. It provides the core of QTP operations. It will be maintained with the highest diligence for security and stability.

- Windows will be employed only where necessary (as forced by marketplace pricing and use of software). By using it on private IP networks only, the security risks are minimized. By deploying the preoxy server, no functionality is lost for the users of NT workstations.

- AIX (SP's), Irix and Mac OS (visualization), Linux (cluster) are deployed on research computing equipment with focus on a limited task, such as number crunching or visualization. By using them on private IP networks only, the maintenance burden of keeping up to date with security patches is minimized.

In general, portable computers and laptops are the responsibility of the owner and are not integrated into the distributed architecture in the sense of being allowed NFS or other privileged access when connected to a QTP network. Visitors can obtain an IP number for their laptops while at QTP using DHCP. DHCP only hands out private IP addresses. Laptop users can use outside services through the proxy server. Disk and printer access from the server is provided through a SAMBA server. For laptops to be supported beyond the minimal network access, they must meet certain security and functionality requirements. Currently Solaris 2.x and NT 4.0 are supported on laptops for members of QTP.

The fact that QTP networks are switched Ethernet makes visitor machines less a problem than they would be in a shared Ethernet environment, but allowing guest hosts does pose a security issue, since internal traffic in QTP networks is not encrypted. For manageability and security reasons, however, supported operating systems must be kept reasonably up-to-date with revision levels and vendor patches. Acceptable use policy does not allow machines running old versions of operating systems (with known flaws) to be connected to the QTP network.

General principles

The architectural implementation just summarized recognizes that the major focus areas for QTP computing are the following:

- **Personal productivity:** (Capacity) This aspect includes the hardware on the desktop, e-mail and web tools, tools to read and write documents (including mathematical equations), tools for making simple graphs and drawings, tools for making spreadsheets, tools to interact with databases, tools for symbolic manipulation (like Maple), and numerical computations (like MatLab).

- **Software development:** (Capacity) QTP develops software to implement the theories it develops. Some members produce software that is intended to be used by others, which requires a higher level of robustness and usability. Others produce software to be used by a few people (sometimes but not always for a relatively short time). In both cases software development tools and compilers are being used and complex code must be debugged. Such work requires hardware resources beyond what is generally needed for desktop productivity. In particular, there can be many relatively short episodes of very intensive numerical work.

- **Production computing:** (Capacity) A group like QTP, that relies on numerical modeling and simulation, needs a large number of CPU cycles and matching RAM and temporary disk space. Typical jobs should not take longer than several days to a week (maximum) to be acceptable.

- **Visualization:** (Capability) Transformation of large and complex datasets to images requires special software tools and special hardware to support them. Two different kinds of visualization should be considered: personal and group. The first kind requires a special graphics workstation, possibly with the ability to generate CD-ROM and video tape output. A group like QTP should have access to at least one and, as demand warrants, several such stations. The second kind requires some large screen display suitable for a seminar room. Such resources could be shared effectively with an entire department like Physics.

- **Extreme computing:** (Capability) Some jobs cannot be done within a reasonable time on what has now become commodity hardware, but require special hardware features and software infrastructure. It has become clear now that the supercomputer of today and for the foreseeable future is a massively parallel distributed memory system with symmetric multiprocessor nodes connected by a fast switch. To use such a machine requires a significant investment to re-engineer existing codes.

The first three items place emphasis on capacity, the other two on capability. The two groups of items must be justified and managed differently. Therefore the strategy can be summarized as a coherent and focussed and flexible environment in which resources for state-of-of-the-art research computing can be explored, deployed and used and in which at the same time a highly functional commodity information technology infrastructure can be provided.

Discussion

Personal productivity and software development

Because of the difficulty in securing funds to keep desktops up-to-date, desktops should be purchased with the primary goal of providing access, i.e. the focus should be on the monitor, which lasts for about 5 years. All services should be provided by a powerful server or set of servers, because incremental funds can be secured to keep the servers up-to-date and make sure they have the required capacity. This model also guarantees that the system is consistent and easy to maintain.

The higher resource demand imposed by the software development mission cannot be met on desktop stations that are affordable. Therefore that task naturally moves to appropriate server(s). The history of QTP shows that with every new desktop purchase, people initially compile and debug on the desktop machines. But within a year such machines no longer are satisfactory for these tasks. Because no funds are available for replacement after a year, the software development moves to the server for the next four years. This cycle is inefficient use of people.

QTP should put software development on one or more servers by design, rather than attempt to keep it on the desktop and repeat the cycle. The same is true to some extent for other activities. That choice implies that QTP will run one or more servers that have the potency of a production node in a non-production mode.

Visualization

With the 2000 IBM SUR award and the 2000 NSF DMR-IMR grant for visualization, Physics and QTP installed a visualization lab in NPB 1213 with nine workstations and one 3d visualization system.

Extreme computing and parallelism

In 1991, QTP was forced to introduce RS/6000 hardware and AIX into the then homogeneous SPARC and Solaris system to get floating point capability up to competitive levels. It was still possible at that time to satisfy the demand for both capability and capacity by adding the RS/6000 cluster as single architecture. In the perod from 1992 through 1998, QTP slowly lost stature as an active state-of-the-art supercomputer user. QTP members still consume large numbers of cycles on Cray T90 machines, but that is no longer super-computing. As an illustrative example, consider that Aces2 was rewritten for vector machines in 1990-1991, which was just before the IBM SP and Cray T3D were announced. Although vectorization has some benefit in super-scalar architectures of modern computer chips, it did not keep Aces2 at the forefront of super-computing. In July of 1996, a 10 node SP system was installed at QTP. This step provided a good opportunity to explore and invest in what is now super-computing technology: parallel programming. However, to this date the SP has been used overwhelmingly to satisfy the demand for computing capacity.

In the Spring of 2000, QTP embarked on new strategy for supercomputing: Taking supercomputers from National Centers as they upgrade their systems. The first system

obtained was Xena phase I, an RS/6000 SP with 134 nodes from the Maui High Performance Computing Center. The second system arrived in the Summer of 2001: Xena phase II is an upgrade of Xena to 176 nodes and of Quant to 61 nodes. This system was retired from the Aeronautical Systems Center (ASC) Major Shared Resource Center (MSRC), located on the Wright-Patterson Air Force Base, Ohio.

No job scales perfectly, even if it is naturally parallel. Even with 99% efficiency, 1% of CPU time is being lost, in comparison with serial jobs. A similar argument can be made at the single processor level and floating point performance. Although a POWER2 at 66 MHz processor is capable of delivering a theoretical speed of over 364 Mflops (4 floating point operations per cycle), one accepts that a sustained rate of 70 MFlops, or 20% of peak, is pretty good.

Therefore, capacity is always used more effectively by running jobs serially. However, using the SP as a single machine adds a capability to run a new kind of job. These are the jobs that cannot be run in a reasonable time in serial mode on any existing serial processor. There is a place for jobs that use one CPU and a large amount of RAM, for example 4 to 8 GB. But some jobs cannot be run in a reasonable time even with large RAM. They require a parallel approach. One should accept that a parallel job running at 90% efficiency is really good and 99% is extraordinary.

Parallelism over a number of processors can be argued to be too expensive in human-power to realize, unless an order of magnitude in reduction of computation time can be expected. To really make use of parallel computing, the software must be designed with massive parallelism as the ultimate goal. Small scale parallelism is useful as an intermediate step and can provide short time benefit, but by itself usually is not worth the effort of re-engineering any software. To reduce the computation from 10 days to 5 by exploiting 2-way paralellism might be useful for some research projects, but the investment usually is too large for such a small pay-off. Nevertheless, the work to make the program run in parallel over two processors for development and debugging purposes, is a necessary, valuable, and often difficult first step towards making it run on tens or hundreds or even thousands of processors. That is the regime where extreme computing starts and where new problems can be solved that cannot be addressed otherwise.

Production computing

The installation of a large SP complex has provided a new opportunity for development and a new source of cycles. The upgrade to Xena phase II will make the large SP complex significantly more useful for all research efforts in QTP.