HPC Subcommittee Report

Review of Computational Hardware Resources Local to PPPL.

December 2005

Paul Henderson

Doug McCune

Steve Baumgartner

<u>Full HPC subcommittee:</u>

S. Baumgartner

R. Budny

S. Ethier

D. Gates

P. Henderson

S. Jardin

D. McCune, chair

J. Menard

D. Mikkelsen

R. Nazikian

E. Valeo.

Where are we now?

Current state of the PPPL HPC environment:

Computational clusters:

Name	Systems	CPUs	CPU	Speed	Year
falcon_1	160	320	Athlon	1.7GHz	2002
falcon_2	20	40	Pentium	1.7GHz	2002
gridn	18	36	Athlon	1.8GHz	2002
kite (infiniband)	24	48	Opteron	2.2GHz	2005
total	222	444			

Large memory computational systems (restricted access):

1) SGI Altix 350 ("mhd", 16 CPUs) 2004
2) SGI Altix Prism ("viz", 6 CPUs) 2004

Computational Infrastructure:

1) pool of 20 general purpose systems (40 CPUs) 2004
(these are the "sunfire" machines).
2) 12 TB storage array 2004
3) 192 TB tape archive 2004
4) updated core servers (NFS, etc.) 2004

Summary: *a robust infrastructure supporting outdated computational clusters.*

Restrictions on use:

In 2004, all systems in the main clusters were combined into a single unrestricted pool of resources, so that no systems were reserved for any specific group. This simplified cluster management and opened the computing environment to more users while not adversely affecting the initial stakeholders.

A handful of cluster-compatible nodes have been purchased with direct funds for specific purposes, which are not available to all users. Two examples of such small machine clusters are ***bigblue*** (10 nodes, 20 CPUs) for NSTX EFIT between-shots reconstructions, and ***swift*** (4 nodes, 8 CPUs) for the Fusion Grid. These are not included in the node counts shown above. The SGI systems were also purchased with direct funds for specific purposes and are not open to all users.

Commitments to PU and sponsors:

The only outstanding commitment to PU is the use (20%) of the SGI Altix system (***mhd***) covered by a Memorandum of Understanding. Paul Henderson contacted Jim Stone and Kai Li at PU to remind them of this available resource, but only Jim has at present replied saying the system might be too small for most of his needs, but that he might like to do 'dome' experiments in the future. Stone will also be involved in the utilization of the new ***IBM BlueGene/L*** system (1024 CPUs) which was recently acquired by PU.

Resources at Princeton University, ORNL, NERSC:

The recently acquired ***jacquard*** cluster at NERSC is online and used by many PPPL researchers, as is the older ***seaborg*** system; a new machine, ***bassi***, is now coming online. The FY-2006 PPPL NERSC allocation is ***2.6 million CPU-hours***; for comparison, a PPPL 400-CPU Linux cluster running year round at capacity would supply ***3.5 million CPU-hours***.

A 64 processor SGI Altix 3700, ***hecate***, has commenced operations on main campus. The PPPL allocation of 25% of this machine (in FY-2006) is targeted to ITER / Burning Plasma modeling by a restricted set of users covered by a direct grant.

ORNL systems (primarily the ***ORNL Cray-X1E***) are also being tested by PPPL staff. Several PPPL groups have applied for time. FY-2006 allocations have yet to be announced; PPPL is expecting a sizable chunk but it will be restricted to one or two select applications.

Scheduling policies at the large supercomputer centers tend to favor large jobs (Ncpu = 256 or more on some systems). ORNL has indicated that only about ten very high end supercomputer applications (from the entire DOE Office of Science) will be supported at their site. In other words, the high end supercomputer centers tend to favor *capability* computing over *capacity* computing. Capacity requirements, i.e. large numbers of smaller jobs, tend to need to be met locally.

PPPL Systems Reliability:

The current ***falcon*** cluster is aging quickly, with researchers reporting up to **50% failure rates** on parallel processing jobs. Non-recoverable system failures have reduced the number of nodes from 192 to 158 in the past year (**an 18% reduction in capacity**). The newer clusters (kite, sunfire) are showing very high reliability and uptime.

Utilization of systems:

In FY2005, we had 96 users of the cluster systems at PPPL. Although CPU slots have usually been available for the next serial job, parallel jobs experienced long queue wait times. Generally, wait times increased sharply as the number of CPUs requested increased—32 processor jobs experienced queue wait times of 1-4 days. Users reported breaking their jobs into smaller pieces (at a cost in labor) in order to get quicker access to machines. The queue wait times show that the clusters are heavily utilized and essentially

saturated. Spot checks at the monitoring sites http://beowulf.pppl.gov/PBSmon.html and http://beowulf.pppl.gov/Ganglia.html typically show 60 to 100% utilization of systems.

The accounting statistics produced by the PBS job scheduler have not proven adequate to show the total system utilization over long periods of time. Although there is a record of the number and duration of jobs correlated to users, the number of CPUs used by each job has not been kept. The shortcomings of this data are recognized; means of improving the situation are actively being pursued. Remedies may include reconfiguration of PBS (if possible) or replacement with a new scheduler (if required). Replacing the scheduler will entail some inconvenience to users (e.g. automated PBS job submission scripts would have to be modified).

In part to make up for the shortcomings of the systems utilization accounting data, a survey was sent to cluster users and responses collected (see Tables 1 and 2 near the end of this report for survey data), and those who responded noted that they run the following applications: GTC, NCSX design, Stellopt, PIES, VMEC, Terpsichore, TRANSP, LPI_particle, m3d, MHD, MHD-AMR, Nozzle, NumSim, 3d_MRI, IDL, FWR2D, gyro, gs2, and NOVA-K.

Table 3 shows PBS-recorded cumulative job times for the top 14 users of the PBS-controlled clusters (about 88% of the total recorded cumulative time). Unfortunately, as mentioned above, an hour long serial job and an hour long 48-processor parallel job each count for just "one hour" in this measure.

Users related to funding:

Please see the usage tables. It is difficult to pin down what users represented what PPPL groups, as most span multiple groups, but generally, Theory, NSTX, CPPG, Graduate Studies, and the Collaboration groups are the main users of the system.

Impact of changing rules of cyber-security regarding access to PPPL:

This is currently unknown. The requirements of HSPD-12 on internal computer access are yet to be defined, as are those for external access. For now, the current SecureID access method meets DOE requirements.

Where do we want to be?

Role of PPPL Computing:

The role of PPPL computing is to provide a reliable infrastructure for general purpose scientific computing, to provide a cluster for small to midrange (1-48CPU) compute jobs in support of our local research, experimentation, and education programs, and to support the development and staging of large jobs (>48 CPU) to ORNL, NERSC, and other sites.

If funding were available, however, our updated infrastructure could support larger scale compute resources that could better complement those at ORNL and NERSC.

Hardware/computing requirements:

A new 400 CPU cluster would support the current load of small to mid-size (1-48 CPU) job runs (allowing ~24 simultaneous 16 CPU jobs) while providing a cushion for incremental growth. Since the current 316 CPU falcon cluster (which runs at 1/3 the speed of currently available hardware) is becoming increasingly unreliable as it ages, it should be replaced.

Currently long running parallel jobs are failing at an increasing rate due to hardware events, forcing users to either resubmit jobs or reduce the number of systems utilized in an attempt to reduce the probability of failure. ***Some researchers report job failure rates of 50%***. This is very wasteful of effort of researchers and support personnel alike, and urgently needs improvement.

Additions to the interactive computing pool of systems should continue as needed. These systems are used for execution of serial jobs, code development, and testing. Currently, 10 ***sunfire*** systems are deployed for general sessions, and have proved much more popular, more cost effective, and more secure than predecessor systems (Sun-5 systems ***orion***, ***taurus***, and the old ***loki***-class alpha linux machines have been retired).

Financing/Funding:

A cluster with 200 dual CPU systems, gigabit cluster interconnect, and 1.5GB of memory per CPU would cost approximately $400,000, about $2000.00 per node. The specific costs and trade-offs (e.g between larger memory/node or more nodes) are rapidly moving targets and also application-dependent; the HPC subcommittee would be interested in recommending specific solutions and participating in a design review process, if a specific budget range could be given.

Infiniband:

24 node (48 cpu) blocks of new cluster nodes can be outfitted with a high speed infiniband interconnect for approximately $30,000, or $1,250 per dual processor node. Assuming $2000.00 base node cost, the incremental cost factor is about 1.63; a break-even cost per unit computation would require codes to run faster by at least this factor to justify the infiniband's cost. The ***kite*** system is outfitted with the infiniband interconnect. The GTC code has run there with a performance enhancement (infiniband vs. Gigabit) of 1.45, somewhat short of break-even. However, Jeff Candy at GA indicated a factor of 4 speedup was attainable with GYRO on a similar 48 processor system. We are trying to verify such performance on ***kite***; more study is needed. Generally, however, results will indicate that infiniband economics are highly application-dependent. Infiniband capabilities are attractive enough that infiniband *compatibility* should be a requirement for new cluster purchases; this requirement can be satisfied with minimal planning and no additional cost.

SGI Purchases:

Direct funds (CSWIM FSP) will likely enable the purchase of an upgrade to the ***mhd*** SGI cluster, from 16p to 32p. These systems are restricted access. These systems are

primarily intended for development of high performance parallel codes which however may run on production systems elsewhere (e.g. ORNL, NERSC). There are times when these systems are not fully utilized, when it might be possible to accommodate one or two additional users. Such access would be at the discretion of the PIs of the projects funding the SGI hardware, and may only be available temporarily. Potential users are cautioned that a significant porting effort can be required to adapt codes to run on this hardware, and that staff for close support of such porting efforts is not available. A smaller SGI cluster, ***viz***, is available for porting and testing codes in an SGI environment.

Infrastructure Requirements

Physical Space:

The current FCC and PPLCC have considerable room for growth, especially as systems are reduced in size (current systems are half the size of the ***falcon*** systems). Physical plant improvements to power distribution, notably a new Uninterrupted Power Supply (UPS), have been installed in FY2005, as have two new A/C units in the PPLCC. A plan is being drafted (J. Anderson, H. Carnevale) to also replace the failed A/C unit in the FCC computer room while also reducing the area of the room being cooled.

Data storage:

An increase of 4TB per year to our storage array, at a cost of $25K/year, is projected to be needed given current growth trends.

Networking requirements:

The PPPL network currently has little spare capacity and would have to be augmented to support new systems. The scope is dependent on the system size, but an expansion would implement a new distributed switching architecture that simplifies system installation. The likely incremental cost is believed to be no more than $15K in FY-2006.

Local and Remote Data Access:

While most data access is local to PPPL, there is significant off-site access to PPPL data, mainly by fusion partners at MIT and GA. In addition, numerous PPPL researchers connect to supercomputing resources at ORNL and NERSC, and transfer data between these sites and PPPL. TRANSP also provides worldwide computational services, and will soon provide a parallelized fast ion package that will increase demand for PPPL TRANSP resources.

Support Personnel:

The UNIX administration staff currently supports both the PPPL HPC clusters and the SGI systems. A new cluster based on commodity hardware imposes minimal additional

load on existing support personnel. This is due to changes in our infrastructure which simplify system deployment (it takes more time to unpack a new system from its box than to add it to our cluster).

However, the incremental cost of adding non-commodity, specialized computing resources (like an SGI running different software than the current SGIs, or, a completely different type of system such as IBM BlueGene) is significant due to specialized hardware, operating system, compilers, and other supporting software requirements.

Usage policies:

We've used strict scheduling policies (where specific systems were reserved for specific groups) in the past with mixed results; the policy generally resulted in under-utilization of systems. Our most effective usage policy is to limit the number of CPUs that can be used by any one user, effectively allowing enough headroom for at least some of the next user's job to always be started.

Still, there has been considerable debate amongst users about optimal scheduling policies. A consensus recommendation is to convene a **regularly scheduled user committee meeting** to review system usage and make recommendations about access and utilization policies. These meetings would function like an engineering design review, in that users could file “chits” that would be specifically answered by a review committee with representation from both the user community and the engineering support staff.

Some limitations of the current PBS scheduler software have been identified. Alternative schedulers and scheduler configurations are available at nominal cost—the real issue being the labor to evaluate alternative solutions, plus a potential transient impact on users should the scheduling software actually change (for example, user scripts that directly utilize PBS would require modification).

Subcommittee Name Change:

HPC or “High Performance Computing” is somewhat of a misnomer of what is being discussed here—namely, midrange capacity scientific computing local to PPPL. A better name for the “HPC planning subcommittee” might be the “PPPL Local Computational Resources planning subcommittee”.

User Survey Results:

Table 1: percentage of users' computational workload on PPPL and elsewhere, as estimated by the users themselves [1].

System:	FALCON	GRIDN	KITE	SUNFIRE	Other or Unspecified
Research Staff					
R. Budny			10%		90%
N. Gorelenkov				90%	10%
S. Hudson				100%	
I. Kaganovich	80%				20%
E. Lazarus[2]	33%				67%
D. Mikkelsen	25%	10%	15%		50%
D. Stotler				95%	5%
R. White	100%				
Grad Students					
Emily Belli	50%				50%
Tom Jenkins	50%			50%	
Wei Liu	100%				
P. Sharma	85%				15%
CPPG					
R. Andre	20%				80%
C. Ludescher	20%				80%
S. Ethier	20%				80%
R. Samtaney		50%	25%		25%
Engineering					
Long-Poe Ku	30%			70%	
TRANSP jobs:	40%			30%	30%

[1] PBS logs also suggested heavy use by graduate students Yoav Avitzour, Adam Sefkow, Peter Norgaard, PPPL researcher Gerrit Kramer, and UCSD collaborating researcher (NSTX project) Alexander Pigarov, none of whom were available to fill out the survey form.
[2] Off-site collaborator.

Table 2: User estimates of typical run requirements; system performance results.

System:	FALCON	GRIDN	KITE	SUNFIRE	SGI/ other
Parallel Job Loads					
#respondents:	5	3	4	1	0
N_cpu	8—32	32	8—32	8	-
Mem/CPU, GB	0.6—2.0	0.6-1.0	0.3—1.0	-	-
Runtime, hours	20—120	48	24—100	25	-
Concurrency*	1—8	1-5	1—6	1—2	-
Queue wait, hours	12—48	0-48	48—100	-	-
System Failure	5—50%	5%	0—5%	0	-
Serial Job Loads					
#respondents	7	0	0	7	0
Memory, GB	0.1—2.0	-	-	0.3—1.0	-
Runtime, hours	3—200	-	-	0.1—200	
Concurrency*	1—20	-	-	1-100	-
Queue wait, hours	0—0.5	-	-	0—2	-
System Failure	1—5%	-	-	0%	-

*Concurrency means, #jobs simultaneously queued and/or executing.

Table 3. PBS statistics—total job runtimes, top 14 users. Statistics *do not reflect N_cpu* of parallel jobs—Nov. 2004 – Oct. 2005.

User	#jobs	Runtime, days	User	#jobs	Runtime, days
Emily Belli	1067	1491	Long-Poe Ku	191	341
Ed Lazarus	10228	1349	Yoav Avitzour	1138	213
Gerrit Kramer	1161	1211	Peter Norgaard	411	199
Wei Liu	254	902	Adam Sefkow	287	186
Prateek Sharma	1194	749	Stuart Hudson	3852	152
David Mikkelsen	582	695	Alexander Pigarov	399	144
Robert Budny	371	536	Pshare*	480	136

*PPPL local TRANSP jobs (mostly NSTX) (not including Fusion Grid).

Recommendations Contained in this Report:

1. Replace Falcon with modern, fast, reliable cluster. At approximately $2000.0/node, $1000.0/CPU, a 200 node cluster would cost about $400,000.
2. Upgrade system utilization statistics and accounting (may entail replacement of PBS scheduler).
3. Hold regularly scheduled user meeting / design review for cluster use and scheduling policies—about once every 6 months.
4. Change the name of the HPC planning subcommittee— "Local Computational Resources planning subcommittee" might be more appropriate.