PPA Computing Survey and Shared Cluster Planning

S. Luitz, SCA Meeting 02/10/12

PPA Baseline Computing Survey

- What are the current and future computing needs of PPA groups and experiments and how could shared computing and storage be provided (funded) considering the ramp-down of BaBar computing?
- Mini-Survey within PPA, quick and dirty, very limited scope
- The basic assumptions:
  - It is shared.
  - CPU access is through batch.
  - Storage is shared and provisioned by amount (not by server).
  - Planning will cover a 5-year period starting with FY12 and take into account evolution or ramp-down of current projects.
  - The size of such a facility would be at the "few $100k / year" (average) for hardware and recharge model cost.

The Survey

- The questions sent out to PPA Scientific Computing Advisory Group (representing PPA groups and experiments):
  - Given the basic assumptions above, please provide:
  - Expected evolution of CPU needs FY12 through FY16 (peak and average) and some indication of the usage pattern. Units of CPU don't really matter as long as they can be reasonably well translated into a number of (contemporary) "cores".
  - Memory requirements corresponding to the CPU use.
  - Expected evolution of the corresponding storage use and some indication of peak/average rates, access patterns (sequential, random).

Reply Summary: PPA Theory Group

- Large-scale Monte-Carlo production with a peak requirement of up to 1000 cores for typically a few days per month, adding up to a few weeks per year. The application is single-threaded and dominated by floating point operations. The memory requirement is 1-2GByte per CPU core. Storage is expected to grow over the next years from approximately 6TB to approximately 20TB. The storage access pattern is random with a high peak rate.
- Continuous tests and validation of Monte-Carlo event generators and the BlackHat library, including user support and production of event samples for LHC experiments. About 100 cores are continuously used. The application can take advantage of multi-threading and MPI and is dominated by floating point operations. The memory requirement is 32GByte/node. Storage needs are about 10TB, the access pattern is random access with low peak rate (I/O during computation)
- Continuous use for data analysis and code development requiring the continuous use of about 100 cores. The applications are multi-threaded and can use MPI or even GPUs and are dominated by floating point operations. The memory requirement is 32GByte/node. The storage estimate is ~4TB, the size not a limiting factor, but the access pattern is random with high a peak rate.

Reply Summary: Computational Cosmology

- Computational Cosmology could take advantage of a computing facility to test codes, simulate and to post-process simulation data sets produced elsewhere. The workload would be very similar to the PPA Theory Group's workload with a mix of traditional batch and HPC processing. A low-latency interconnect for MPI processing, scheduling and reservation policies that would allow MPI processing and high-bandwidth temporary storage (parallel file system, a few tens of TB) would be highly desirable. Memory requirement are >= 4GByte/core. Computational Cosmology could saturate a system up to 1000s of cores for extended times, so their use of a shared facility would not be driven by demand but limited by availability and allocation policy.

Reply Summary: DES

- DES has three different types of workloads:
  - Full cosmological simulations, to generate mock catalogs. This is MPI computing requiring a large number of cores, lots of memory, and high bandwidth. Whether this is done at SLAC or offsite on other resources is still to be decided. As a highly DES-specific resource requirement it is not within the scope of this document. The post processing of the generated data would be done at SLAC and requires a cluster with a low-latency interconnect, but the jobs would only require up to 200 cores.
  - Cluster finding (on the data from 1. above). This will be driven largely by a soon-to-arrive new staff member and is generally just CPU and memory bound.
  - Cosmological constraint estimation. This is embarrassingly parallel and largely CPU bound.
- Overall, the components of DES computing that are relevant for this document would continuously saturate about 260 cores. CPU requirements are expected to be fairly constant over the next 5 years. Memory requirements are at least 4GB per core and at least half of the machines should have a low-latency interconnect. In addition to approximately 50TByte of short-term working space, long-term disk storage capacity is expected to grow at a rate of about 100TByte/year over the next 5 years.

Reply Summary: FGST

Fermi Gamma-Ray Space Telescope (FGST)

FGST has a fair share CPU allocation of 2400 cores for pipeline processing, event reconstruction, reprocessing, and Monte Carlo simulation. However, FGST has been taking advantage of the general SLAC scientific computing infrastructure to absorb their peak loads for reprocessing.

The CPU need for continuous processing is expected to remain constant at 250 cores, peak CPU (in excess of the continuous need) demands are expected to grow from 900 cores in FY 2010 to 2800 cores in FY 2016. Peak CPU is needed for about 6 months a year (reprocessing and Monte Carlo simulations), some of it may be provided by other collaborators. FGST memory requirements are 3-4 GByte/core

Funding for FGST storage remains dedicated and is not in scope for this proposal.

Reply Summaryies: EXO, CDMS, SuperB

- **EXO**
  - Current use is at about 100 CPU cores with short peaks of up to 1000 cores (provided by the shared batch system). No projections of future CPU needs (including EXO-200) are currently available. EXO has currently about 130TByte of dedicated disk space, expected growth is at about 30TByte per year. Experience shows that I/O performance needs to be managed so a shared storage model would need to be tested carefully.
- **CDMS**
  - Currently CDMS uses the batch farm for MC production only with an allocation of 100 cores. Jobs do not need very much memory (less than 1GByte). Peak usage is more than 100 cores but currently unknown. The 100-core allocation would likely be even adequate for possible data processing at SLAC. Data storage requirements for experimental data would be in the tens of TByte range.
- **SuperB**
  - SuperB could take advantage of opportunistic cores during large-scale simulation production campaigns. The workload is Grid/traditional batch, memory requirements are 2-3 GByte / core. There is a small base workload (10s of cores) for local development and analysis. Peak core use would be ~2000 cores for 2-3 weeks a few times a year. SuperB has 15TByte existing shared storage (NFS), the shared storage need would increase to about 100TByte in FY16. Potential development of a SuperB Tier-1 is out of scope

CPU Base and Peak Loads

Group/Experiment/ Project	Cores Base Load [FY 12 Cores]						Peak Cores (in addition to base cores)					
	FY12	FY13	FY14	FY15	FY16	Base Notes	Peak Uage Pattern	FY12	FY13	FY14	FY15	FY16
PPA Theory	200	200	200	200	200		a few days per month adding up to a few (5)? weeks per year	1000	1000	1000	1000	1000
DES	256	256	256	256	256		not in scope	>1000	>1000	>1000	>1000	
LSST database tests	0	0	0	0	0		need to reserve a large number of machines (250), 2-4 times a year.	"4000"	"5600"	"8000"	"11200"	"16000"
LSST data challenge	0	0	0	0	0		up to a week, 2-4 times a year	100	100	100	100	100
LSST simulation production	370	580	780	780	780		4 months/year	370	580	780	780	780
Computational Cosmology	100	100	100	100	100	Guess	not in scope	>1000	>1000	>1000	>1000	>1000
EXO	100	100	100	100	100		a few days per year, 3-4 times	1000	1000	1000	1000	1000
CDMS	100	100	100	100	100		none	0	0	0	0	0
FGST	0	0	0	0	0	out of scope	6 months/year	900	900	2300	2300	2800
SuperB	10	20	30	40	50		2-3 weeks/year 3-4 times, Grid workload	2000	2000	2000	2000	2000
Total	**1136**	**1356**	**1566**	**1576**	**1586**			**5370**	**5580**	**7180**	**7180**	**7680**

	FY12	FY13	FY14	FY15	FY16
Total Base+Sum of Avg Peak	**2316**	**2606**	**3583**	**3593**	**3853**
Total Base+Sum of Peak	**6506**	**6936**	**8746**	**8756**	**9266**

SLAC NATIONAL ACCELERATOR LABORATORY

Averaged Peak Loads

- To see the scale of our peaks, let's look at the peak CPU if it could be smoothed out over the whole year

Group/Experiment/ Project	Averaged Peak Cores				
	FY12	FY13	FY14	FY15	FY16
PPA Theory	100	100	100	100	100
DES					
LSST database tests					
LSST data challenge	7	7	7	7	7
LSST simulation production	123	193	260	260	260
Computational Cosmology					
EXO	71	71	71	71	71
CDMS					
FGST	450	450	1150	1150	1400
SuperB	429	429	429	429	429
0					
Total	**1180**	**1250**	**2017**	**2017**	**2267**

Storage

Group/Experiment/Project	Storage (TB)					
	FY12	FY13	FY14	FY15	FY16	Storage Notes
PPA Theory	21	25	30	35	35	
DES	50	150	250	350	450	
LSST database tests	0	0	0	0	0	local storage only
LSST data challenge	5	15	20	20	20	guessed
LSST simulation production	5	15	20	20	20	guessed
Computational Cosmology	25	50	50	50	50	"a few tens of TB"
EXO	0	30	60	90	120	"130TB dedicated storage exist, only additional storage at 30TB/year"
CDMS	5	30	30	30	30	Experimental data a few tens of TB
FGST	0	0	0	0	0	FGST has dedicated storage
SuperB	0	20	40	75	100	replacing current shared storage
Total	**111**	**335**	**500**	**670**	**825**	

SLAC NATIONAL ACCELERATOR LABORATORY

First (and obvious) Observations

- >=4GByte/core RAM are a requirement for a common platform
- There are clear opportunities for sharing and to absorb peak loads in a shared facility.
  - Devil will be in the (scheduling) detail.
- Low-latency interconnects and fast temporary storage desirable for the “Theory” applications
- Sufficient amount of local disk required for at least one application
- Initial storage requirements seem fairly moderate

Proposed Approach: Compute

- Purchase a cluster at a rate of N (N determined by funding) nodes per year
  - » Model: Per –"node" cost constant, CPU power per server doubles every two years
- Replace nodes in their 5th year, at this point keep cluster size constant
- Initial configuration (high-level)
  - » Dual-hex core CPU, >=48GByte RAM, at least 2 2-3TByte local SATA disks per server, 2 x 1GBit ethernet link, QDR FC
  - » Include 50 Tbyte of shared "local" storage to the cluster
- Develop a model how to share the different workloads on the cluster

Proposed Approach: Storage

- Transition from a model where groups purchase disks and storage servers to a model of “shared” storage that is billed by size (and probably performance tiers).
- Access through parallel file system
  - » Will need to choose
- While namespace is shared, underlying devices may or may not be shared (c.f. AFS)
  - » Naïve approach to performance tiers:
    - Shared disks (raid set) and/or server
    - Dedicated disks (raid set) and/or server
    - Custom (SSDs, faster servers, additional replicas, etc.)
- 1st step: Purchase an initial test bed configuration in FY12 that would also satisfy the initial requirements in this model and would be extensible / scalable later
  - » Storage testing with synthetic workloads is difficult
- Transition to a storage rent model in FY13
  - » Exact financial models to support this are as of yet unknown

Model for a $500k/year Cluster

CPU base requirements met in FY13

Base + >80% of averaged peak requirements met in FY15

Base + >80% peak met in FY16

100% of Storage

PPA Compute Cluster	All cost in k$							
Fiscal Year	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	Notes
# servers purchased	**60**	**60**	**60**	**60**	**60**	**60**	**60**	
# 5-year old servers retired	0	0	0	0	0	60	60	
# servers in cluster	**60**	**120**	**180**	**240**	**300**	**300**	**300**	
CPU added to the cluster [FY12 cores]	720	1018	1440	2036	2880	4073	5760	
CPU retired from cluster [FY12 cores]	0	0	0	0	0	720	1018	
Total cluster CPU	**720**	**1738**	**3178**	**5215**	**8095**	**11448**	**16189**	
Cluster local storage [TB]	**50**	**100**	**150**	**200**	**250**	**300**	**350**	
Local storage cost per TB	0.60	0.42	0.30	0.21	0.15	0.11	0.08	
Storage hardware cost for cluster	30	21	15	11	8	5	4	
Purchase cost (servers + IB HBA/cable + network)	354	354	354	354	354	354	354	
Infiniband infrastructure	30	30	30	30	30	0	0	
Total recharge cost for cluster	13	26	39	51	64	64	64	recharge cost for full FY
Total cluster cost	**427**	**431**	**438**	**446**	**456**	**424**	**422**	
Storage								
Storage as a service, cost per Tbyte/ year	0	0	0	0	0	0	0	
Shared Storage Testbed investment	75	0	0	0	0	0	0	one-time
Testbed storage provided	120	120	120	120	0	0	0	assume test bed retires in FY15
Storage as a service, TB required	111	335	500	670	825	985	1135	extrapolated for FY17+18
Storage as a service, annual cost	0	23	29	29	31	26	21	
Total Storage cost	**75**	**23**	**29**	**29**	**31**	**26**	**21**	
Grand total cost for FY	**502**	**454**	**466**	**475**	**487**	**450**	**443**	

Notes:

Large uncertainties in projections

Assuming continuation of a usable CPU/ cost ratio doubling every 2 years

Actual CPU needs often expand or shrink to fit into available resources

SLAC NATIONAL ACCELERATOR LABORATORY