

Petascale Platforms: Expect the Unexpected

**Presented to Petascale Systems Integration into
Large Scale Facilities Workshop
Hotel Nikko, San Francisco, California
May 15-16, 2007**

**Dr. Mark K. Seager
Lawrence Livermore National Laboratory**

UCRL-PRES-TBD

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.



Outline



- ✦ Petascale Facilities Challenges
 - ◆ Your mantra is “maximize flexibility”
 - ◆ Power density is going up, way up
- ✦ Platform and usage model drive infrastructure challenges
 - ◆ Archive, hold my bits forever
 - ◆ NAS, what you want IOPS too?
 - ◆ SAN is not NAS spelled backwards
 - ◆ Scalable parallel file system
 - ◆ Visualization
- ✦ Platforms Challenges
 - ◆ Shifting machine balance means FLOP/s are free
 - ◆ Development is managing risk
 - ◆ Integration is making lemonade out of lemons
- ✦ Note: In this talk we ignore the hardest problem – delivering petascale science





NNSA's "Challenge" computers require a dedicated petascale computing facility and staff



- Integration and support at this scale will require an experienced, specialized and highly-skilled staff
- Even though these systems and their infrastructure are assembled from commercially available components, scale of integration is unique
- Close collaboration with vendor partner is essential for success
- Highly specialized customer service organization required to enable end-user productivity

Number of FTEs required to procure, site, integrate and serve a major "Challenge" environment over and above personnel needed to operate "Workhorse" environment

systems	networking + security	storage	customer service	tools and performance	visualization	management and procurement	operations	facilities	total	Cost (\$M)
16	6	9	10	11	8	6	8	4	78	23.4



Customer Service is part of the culture and maximizes return on the investment in the systems



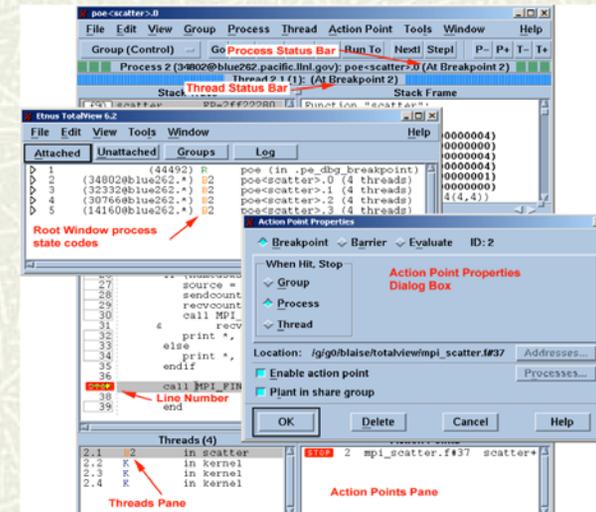
HPC Help desk available 9x5, Operations—24x7

- ◆ Over 2400 active customers
 - 700 remote customers include:
 - ASCI Alliances, Tri-Lab customers, Collaborators, Institutional users
 - 97% of 11800 Remedy tickets in 2004 closed
 - 106651 Remedy tickets since 1996
 - Dedicated Application Time (DAT) assistance
 - On all parallel machines for Priority work



Outstanding web presence for users

- ◆ Online user manuals
- ◆ Status and monitoring of HPC systems
- ◆ Web tutorials





Petascale systems require a significant investment in facilities



Computing complex consists of TSF (B453), B451, B439 and B115

- ◆ \$91M TSF facility completed a year ahead of schedule, operations began December 2004
- ◆ TSF
 - 47,500 ft² computer rooms
 - 253,000 ft² total
 - 20.4MW expandable to 45MW (run + cool)
- ◆ LC Complex (includes TSF)
 - 75,000 ft²
 - 29 MW expandable to 53 MW (run + cool)
- ◆ Can site three Challenge systems simultaneously
 - White (in service)
 - Purple 100TF final delivery: August, 2005
 - BG/L final 32 racks delivery: June 2005
- ◆ FY06 LC Program operating budget \$78M



32 racks of BG/L under integration



Terascale Simulation Facility

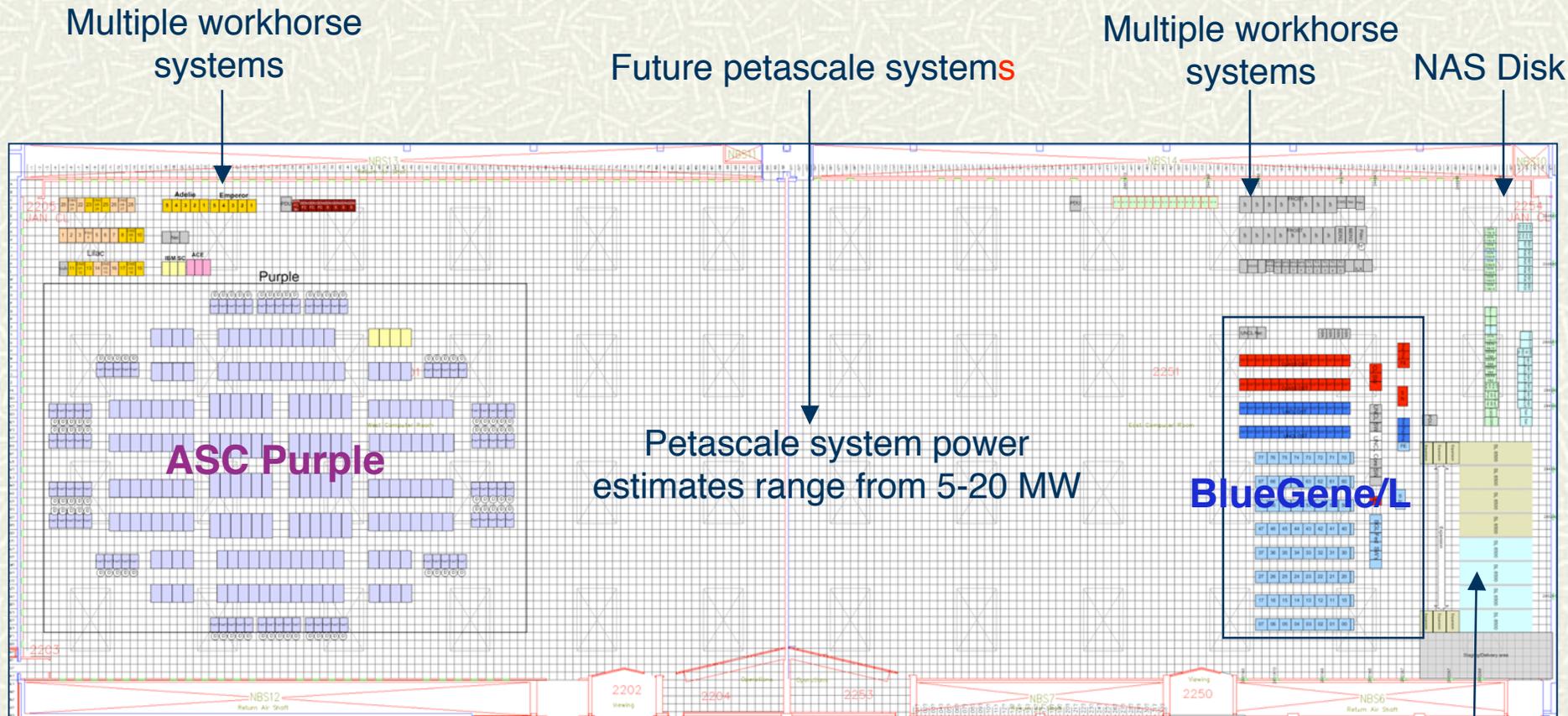
—Security Level Limited, Q

—Hazard Level: Standard Industrial

LC Program Facility Costs (\$K)	FY05	FY06 (est)
Plant Engineering	486.00	500.00
Machine Site Prep	3,000.00	500.00
Move	834.00	-
Power	5,610.00	11,500.00
OFC	6,322.00	8,000.00
	16,252.00	20,500.00



TSF highly efficient machine room design allows for low overhead cooling (35%)



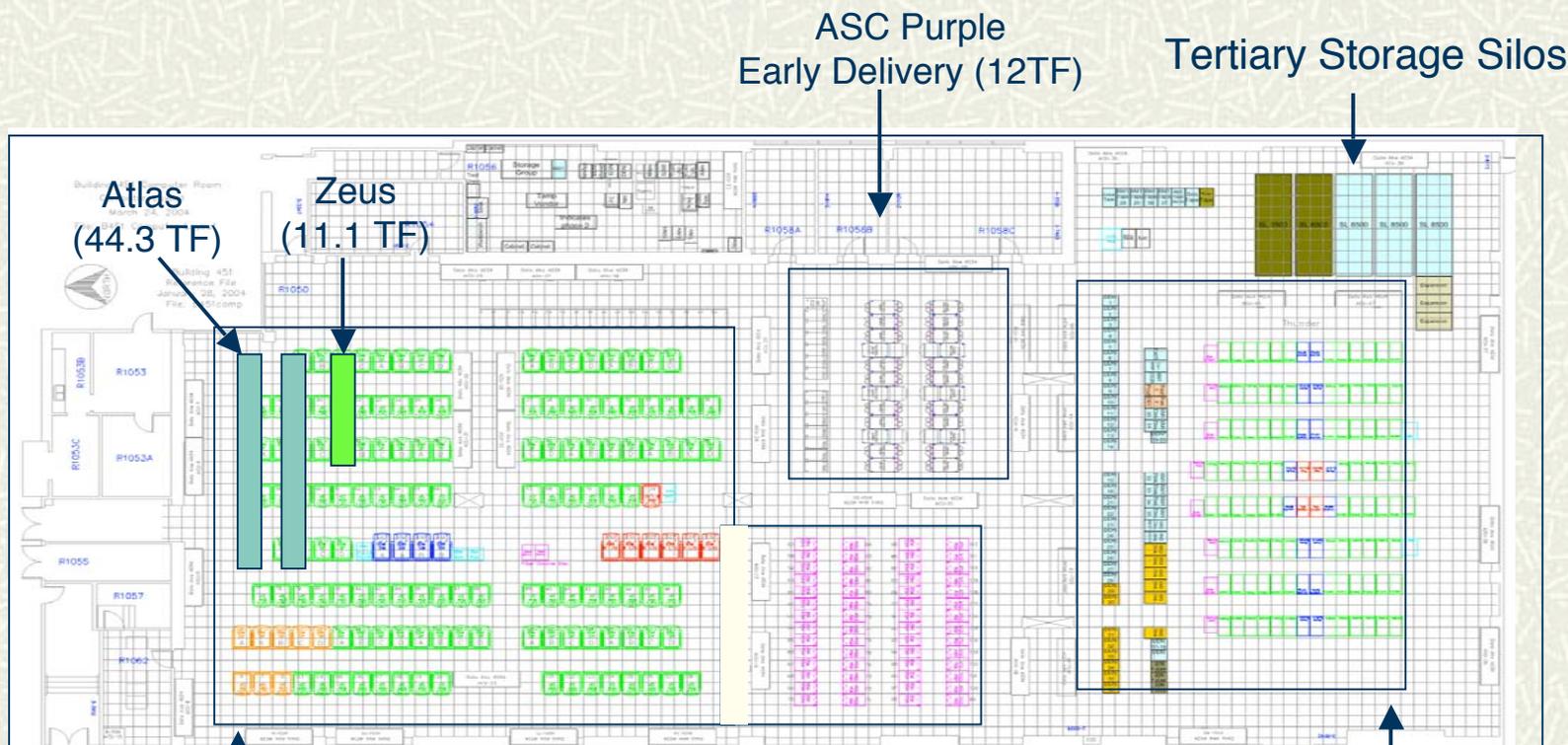
Petascale system power estimates range from 5-20 MW

This compute capability combined with 12.5 (going to 30) MW and 47,500 ft² available for computers is unique national asset

Tertiary Storage Silos



B451, next door, provides significant additional capacity - currently fields 90.4 TF alone, more than at all but a handful of sites across the world



This building provides supplementary Infrastructure including 6+ MW and 20,000 ft² of high quality space. There are three other buildings siting systems and support equipment as well



Facilities Requirements



- ✚ Two word summary – maximum flexibility
 - ◆ Clear span
 - ◆ Air and water cooling options
 - ◆ Multiple machine deployments with 1.5:1 size ratio
 - ◆ Ability to field racks with 208V 3Phase or 480V 3Phase
 - ◆ >250lbs/ft² floor loading
- ✚ Power density increasing
 - ◆ 3.9MW/19Kft² for 10TF/s 2002
 - ◆ 12.5MW/48Kft² for 100TF/s 2005
 - ◆ 30MW/48Kft² for sustained PF/s 2008-2011



Purple Procurement Activity



- ✚ Purple budget line had multiple contracts
 - ◆ IBM Purple
 - ◆ Cisco for SAN, DDN for Lustre OSS
- ✚ Purple contract had two SOWs
 - ◆ pEDTV, Purple
 - ◆ ALC, BlueGene/L



Purple Active Risk Management



- ✦ Active program management of all aspects simultaneously
 - ◆ Weekly tactical calls
 - ◆ Monthly technical deep dives
 - ◆ Quarterly executive focus meetings
- ✦ Risk mitigation plan identified risks
 - ◆ Probability of occurrence
 - ◆ Impact to program
 - ◆ Mitigation plan with decision dates and owners
 - ◆ Required IBM executive approval and commitment



Most Purple risks did eventuate



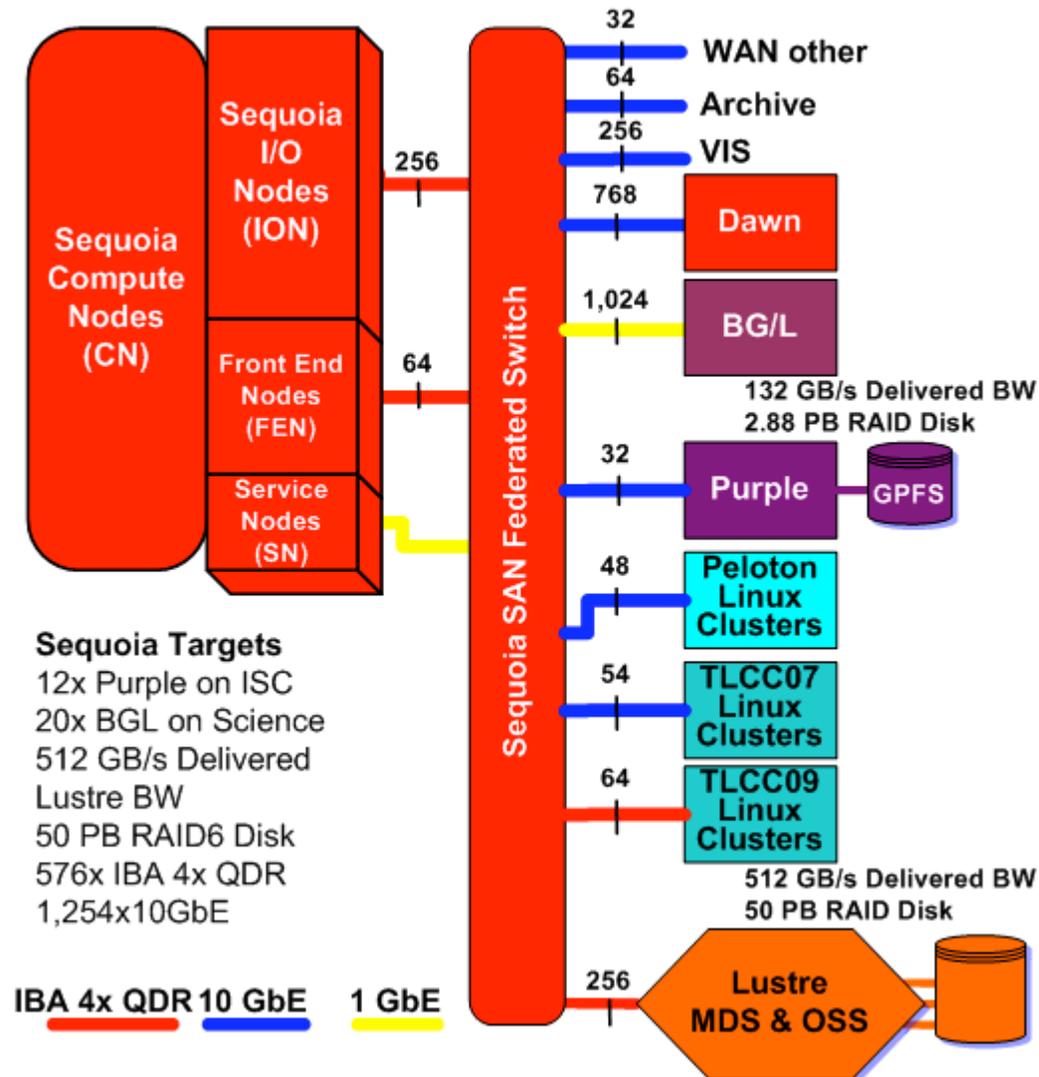
- + EDTV
 - ◆ Software scaling problems
 - ◆ Federation adapter and performance problems
- + ASC out year budget risk
 - ◆ Lead to fundamental Purple 100T change → 1,536 SMP (8xPower5) and three stage Federation
- + Scaling Purple took 2-3 months longer
 - ◆ IBM responded by reducing time to build Purple at LLNL
 - Required delivery of second Federation switch
 - Would not have happened without detailed risk plan with committed risk mitigation strategies and decision dates
 - ◆ Shipped 1/6th of Purple early, so program could have use of machine while additional scaling was completed
- + Purple integration issues were resolved while running ASC Workload at 95-98% utilization doing RRW design
 - ◆ MTBF improved from 12.3 hrs to 80 hrs
 - ◆ Replace entire IO subsystem



A petascale simulation environment drives huge infrastructure budgets



ASC Sequoia Simulation Environment Lawrence Livermore National Laboratory 2010/11

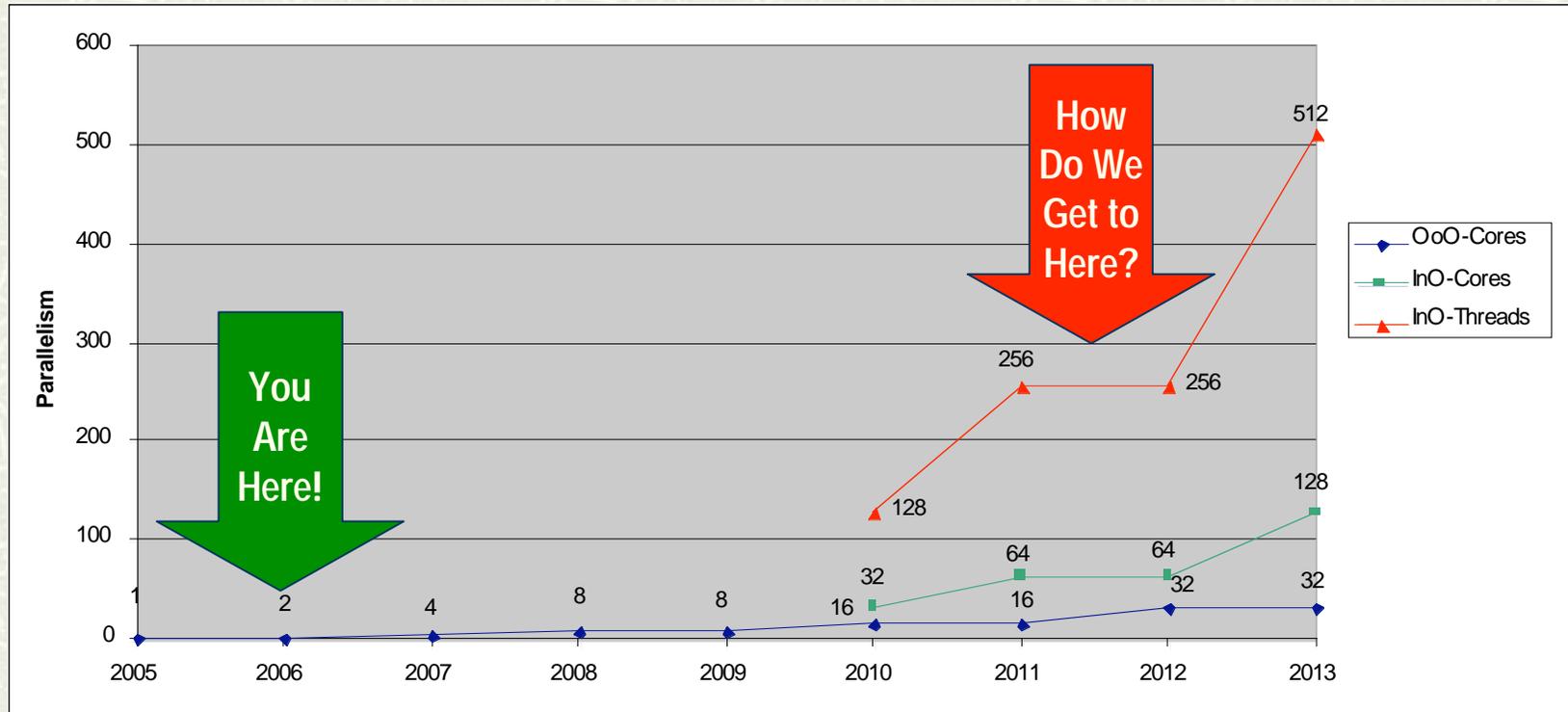


- ✦ \$175M Platform
- ✦ \$25-40M Parallel file system
- ✦ \$25-35M Archive
- ✦ \$5M SAN
- ✦ \$3-5M NAS
- ✦ \$1.5M Visualization

- ✦ Infrastructure is \$60-87M or 35-50% of platform costs



How many cores are you coding for?



Microprocessor parallelism will increase exponentially in the next decade



How much parallelism will be required to sustain petaFLOP/s in 2011?



- ✚ Hypothetical low power machines will feature 1.6M to 6.6M way parallelism
 - ◆ 32-64 cores per processor and up to 2-4 threads per core
 - ◆ Assume 1 socket nodes and 25.6K nodes
- ✚ Hypothetical Intel terascale chip petascale system yields 1.5M way parallelism
 - ◆ 80 cores per processor
 - ◆ Assume 4 socket nodes and 4,608 nodes (32 SU of 144 nodes with IBA)
- ✚ Holy cow, this is about 12-48x BlueGene/L!



Multicore processors have non-intuitive impact on other machine characteristics

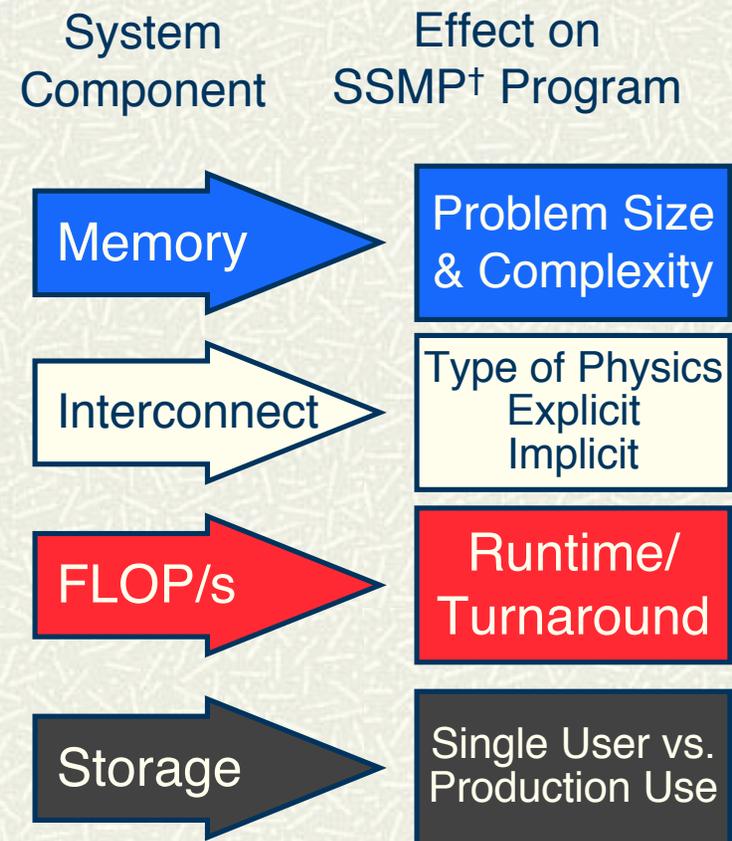


+ Memory considerations drive a change from current practice

- ◆ ASC applications require >1GB/MPI task
- ◆ If we map MPI tasks directly to cores, then platform memory would not be affordable nor practical

+ Interconnect considerations also drive a change

- ◆ ASC applications require >2 million messages/s/MPI task
- ◆ If we map MPI tasks directly to cores, then the resulting requirements are not achievable





Sequoia Procurement Strategy



- + Two major deliverables
 - ◆ Petascale Scaling “Dawn” Platform in 2008
 - ◆ Petascale Weapons “Sequoia” Platform in 2011
- + Use lessons learned from ASC and Linux Cluster Procurements
 - ◆ Cost performance model driving Linux to petascale
 - ◆ Leverage best-of-breed for platform, file system, SAN and storage
 - Separate just in time procurements for each
 - Sequoia budget covers all components
 - ◆ Major Sequoia procurement is for long term platform partnership over two generations and five year development span and five year production span
 - ◆ Risk reduction built into overall strategy from day-one
- + Drive procurement with single mandatory
 - ◆ Delivered petascale performance on marquee benchmarks
 - Few in number (2-5)
 - Representative of critical weapons packages
 - Focus partnership on improving performance of these key benchmarks
 - ◆ Timescale, budget, technical details as target requirements
 - ◆ Focus on TCO factors as well as technical attributes in evaluation



Pro-Active Risk Management Strategy



- + Require detailed project management plan
 - ◆ Manage program similar to White and Purple
 - Weekly tactical telecons
 - Monthly technical reviews/interactions
 - Quarterly tri-Laboratory executive reviews
 - ◆ Detailed WBS, Gantt chart
 - ◆ Risk reduction plan
 - Categorize risks → Impact and probability of occurrence
 - Plan identifies someone responsible, decision dates and mitigation plan
- + Manage high-level risks
 - ◆ Partnership language
 - ◆ Sequoia deliverable has design targets that are later firmed up into requirements
 - ◆ Evaluate prototype hardware/software
 - ◆ GO/NOGO build decision point
 - ◆ Few clearly defined exit points for both parties
- + How do we deal with $O(1M)$ parallelism?
 - ◆ Develop and scale algorithms and applications with platforms
 - ◆ Dawn critical for applications scaling and technology evaluation
 - Will have programmatic impact
 - ◆ Major Sequoia milestones before build and after acceptance



Summary



- ✦ Petascale systems requirements require flexible approach to facilities
- ✦ Infrastructure costs are a major fraction of the platform costs
- ✦ Building, integrating and getting useful work out of petascale systems will be a daunting challenge
 - ◆ Requires active risk management
 - ◆ At some point, you have to make lemonade out of lemons