

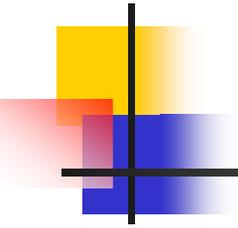
Hardware Plans for FY08-FY09

Don Holmgren
Fermilab

2007 Annual Review

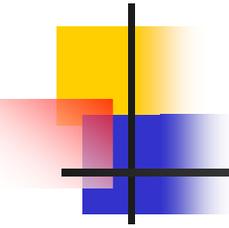
Thomas Jefferson National Accelerator Facility

May 14-15, 2007



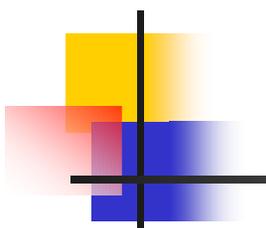
Outline

- Advantages to Combining FY08/FY09 Acquisitions
- Computer Room Planning and Purchasing Details
- Schedule
- Cluster Performance



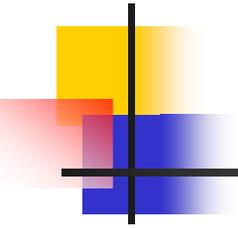
Summary

- Project plans as stated in the FY08 OMB-300:
 - Deploy a cluster, capacity 4.2 Tflops, in FY08, release to production June 30, HW budget: \$1,630K (\$1,460K after G&A, storage, QCDOC)
 - Deploy a cluster, capacity 3.0 Tflops, in FY09, release to production June 30, HW budget: \$798K (\$630K after G&A, storage, QCDOC)
- Proposal:
 - As proposed at last year's review, combine FY08 and FY09 purchases into one purchase
 - Use one RFP and contract, with two deliveries
 - Acquired system to be deployed at Fermilab
 - Release to production: January 2009 (midpoint of OMB-300 FY08 and FY09 deployment dates)
 - Estimated performance: 6.2 TFlops (conservative goal)



Advantages to a Combined FY08/FY09 Acquisition

- The budget profile allows for a large purchase (\$1.46M) in FY08, and a smaller purchase (\$0.63M) in FY09
- There are scientific advantages to combining the smaller FY09 acquisition with that from 2008 for a single larger system (see next slides)
- This FY08/FY09 system will be deployed at Fermilab
 - JLab would not be able to take a large FY08 system without base funding profile changes for the required infrastructure
- At the May 2006 DOE Project Progress Review, the review committee endorsed this approach

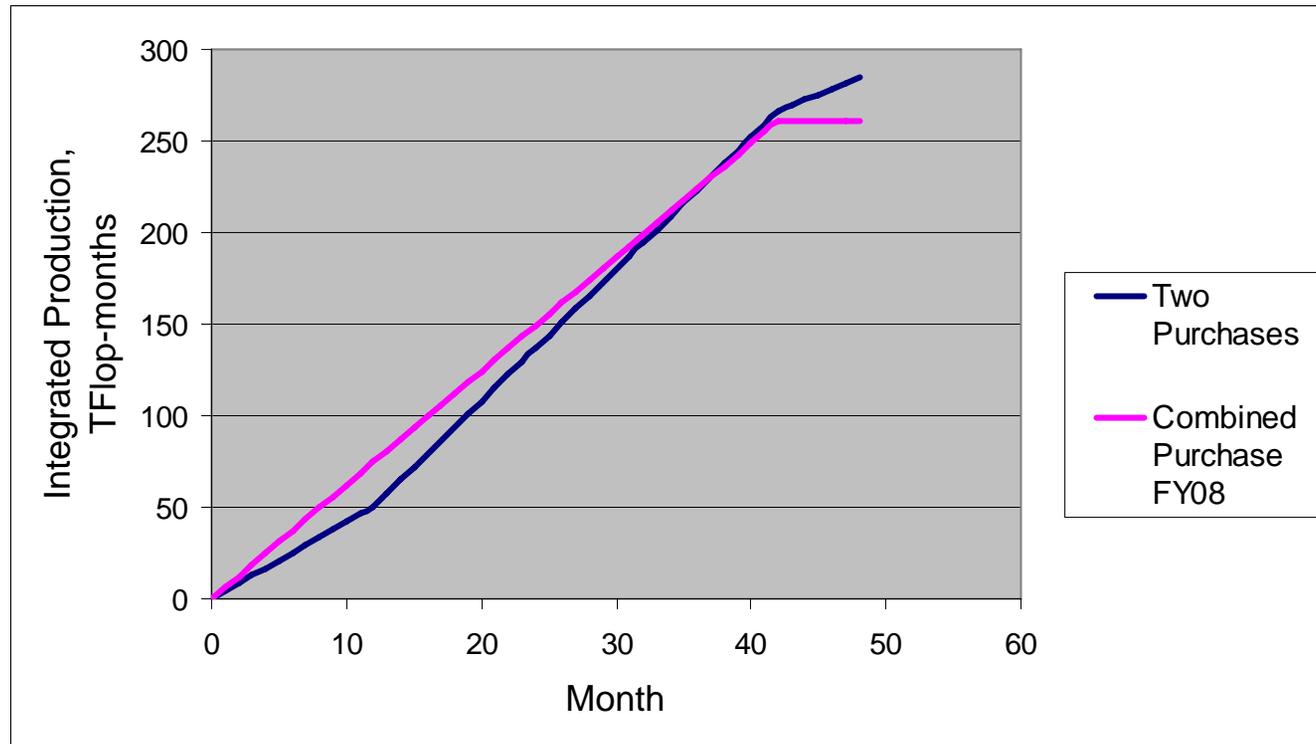


Advantages, cont'd

- Disadvantages of a small FY09 purchase:
 - Because of Moore's Law, we would expect faster hardware to be available in FY09
 - However, faster hardware could not be integrated with an FY08 system in the sense of jobs spanning both sets of hardware
 - A larger gauge configuration generation machine would result from a combined FY08/FY09 purchase
 - Integrated physics production would be greater through the first 36 months of operation
 - Procurement requires manpower, both on-project and in-kind

Advantages, cont'd

Compares: FY08 4.2 Tflop + FY09 3.0 Tflop
with combined in FY08 $4.2 + 2.0 = 6.2$ Tflops
(assumes retirement after 3.5 years, 21-month halving time)



Advantages, cont'd

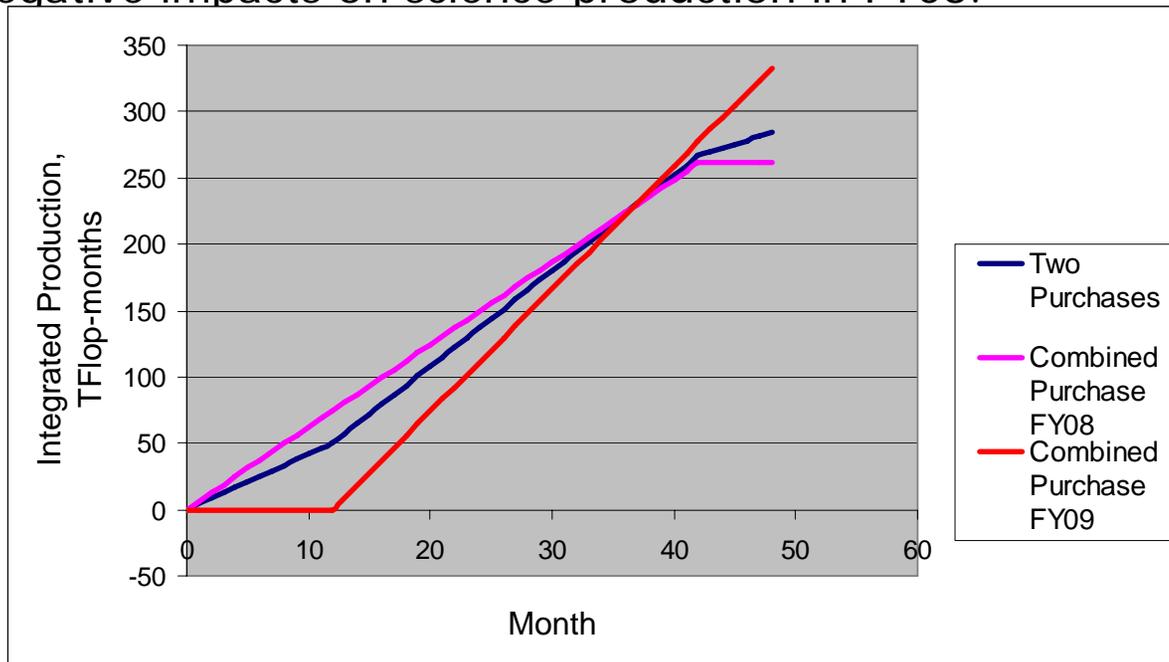
Compares: FY08 4.2 Tflop + FY09 3.0 Tflop

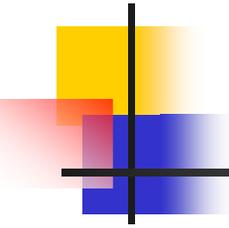
with combined in FY08 $4.2 + 2.0 = 6.2$ Tflop

with combined in FY09 $6.2 + 3.0 = 9.2$ Tflop

(assumes retirement after 3.5 years, 21-month halving time)

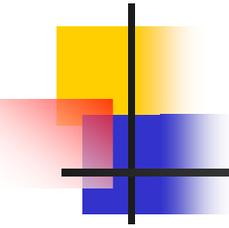
Note – deploying the combined machine in FY09 rather than FY08 would have significant negative impacts on science production in FY08.





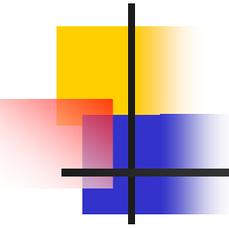
Computer Room Planning

- Fermilab will house the FY08/FY09 machine – “J-Psi” - in a new computer room, “GCC-C”
 - GCC-C will house computers from a variety of Fermilab projects, including LQCD, US-CMS, accelerator modeling, and Fermigrid / Open Science Grid
 - GCC-C build-out (physical building already exists) will begin in FY08 and will take 4-6 months to complete
 - Requisitions are in process for long lead-time items for GCC-C
 - Bids will be in hand by the start of FY08
 - If a C.R. delays project start until March 2008, a 6-month construction interval plus 1 month of float allows beneficial occupancy for LQCD in October 2008



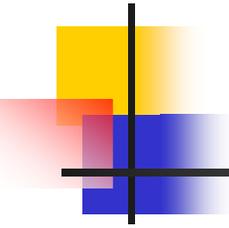
Computer Room Planning

- GCC-C physical constraints:
 - 48U racks
 - 2-foot raised floor plenum
 - 15 KW of power and cooling available per rack
 - 120V or 208V (our choice) overhead power
 - Overhead cable trays
- Integration with Fermilab clusters
 - Storage (dCache, NFS servers) will remain in the LCC computing building that houses the Kaon, Pion, and QCD clusters
 - 10 gigE or Infiniband range expander will be used to couple the J-Psi and Pion Infiniband fabrics (for file I/O only, not for MPI traffic)



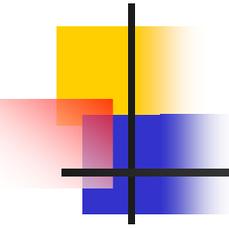
Purchasing Details

- An option specified in the RFP and implemented in the subcontract used in FY08 for the purchase of “J-Psi” will allow us to use FY09 project funds as soon as they are available to buy additional identical hardware from the vendor
- If funds can be released in October to exercise this option, delivery of the additional FY09 systems will occur in December
- The combined FY08/FY09 hardware will be released to production in January
 - “Friendly User” production on the hardware delivered in October will begin in mid-November
 - A similar two-phased delivery was used on the Kaon cluster in 2006



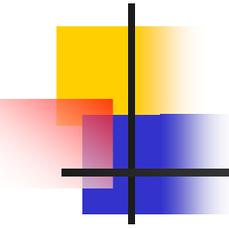
Schedule – 2007/2008

- Summer 2007: Evaluate Intel “Penryn” processor family and “Seaburg” chipsets, and AMD “Barcelona” processor family
- Fall 2007: Test Intel “Nehalem” processors and “Tylersburg” chipsets
- Feb 15, 2008: Preliminary Design Document complete
- Mar 15: RFI released to vendors
- Apr 15: RFI response evaluation complete
- May 15: RFP released to vendors
- June 15: RFP award recommendation complete
- July 1: Contract awarded (commit FY08 funds)
- Aug 1: Approval of sample unit
- Aug 15: Approval of 1st rack
- Oct 1: Delivery of remaining equipment



Schedule - 2009

- Oct 15: Exercise of purchase option (commit FY09 funds)
- Nov 15: "Friendly User" production begins on FY08 portion of J-Psi
- Dec 15: Delivery of FY09 equipment
- Jan 1: Release to production of FY08 portion of J-Psi cluster
- Jan 15: "Friendly User" production begins on FY09 portion of J-Psi
- Jan 31: Release to production of full J-Psi cluster

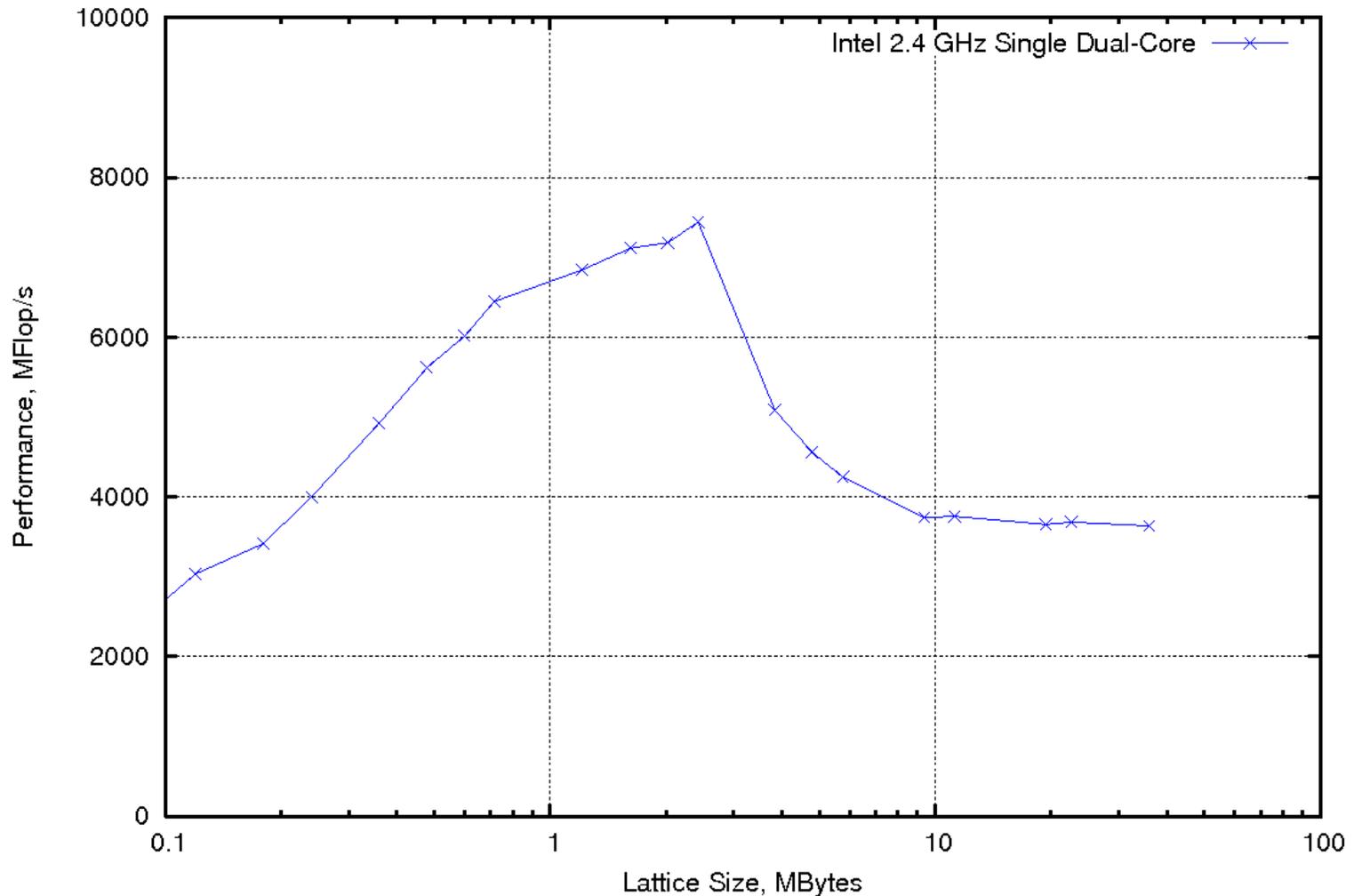


Performance

- Processor candidates:
 - Quad-core Opteron: next spin should include larger L3 caches
 - Quad-core Pentium and Xeon
 - Single socket DDR3 (next gen), or dual socket Xeon with FB-DIMMs (next gen), or dual socket Xeon with DDR3 (next-next gen)
 - Next generation chipsets and CPU's will be available for purchase August 2007; LQCD will benchmark early releases in June/July
 - Next-next generation chipsets and CPU's will be available by late Fall for benchmarking, and Intel is confident of delivery in October 2008
 - Anticipate better memory bandwidth (1600 MHz memory bus, improved FB-DIMM chipset, aggressive DDR3 "CSI") and L3 caches
- Network:
 - Infiniband, DDR or QDR, using anticipated 32-port crossbar switches
 - Next-next generation could use PCI-Express2 (better latency)

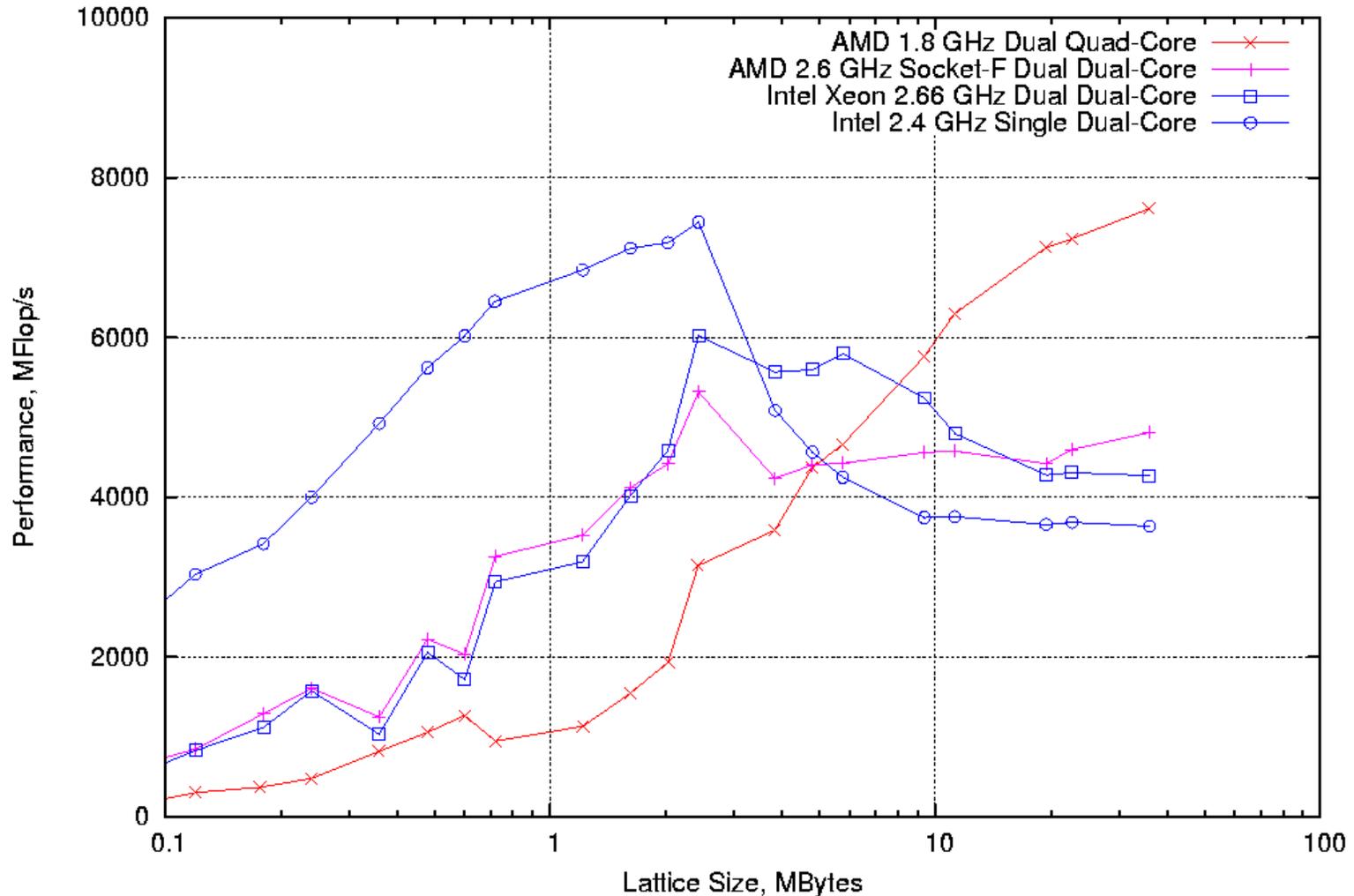
Performance on LQCD Codes

Performance of MILC Improved Staggered Inverter



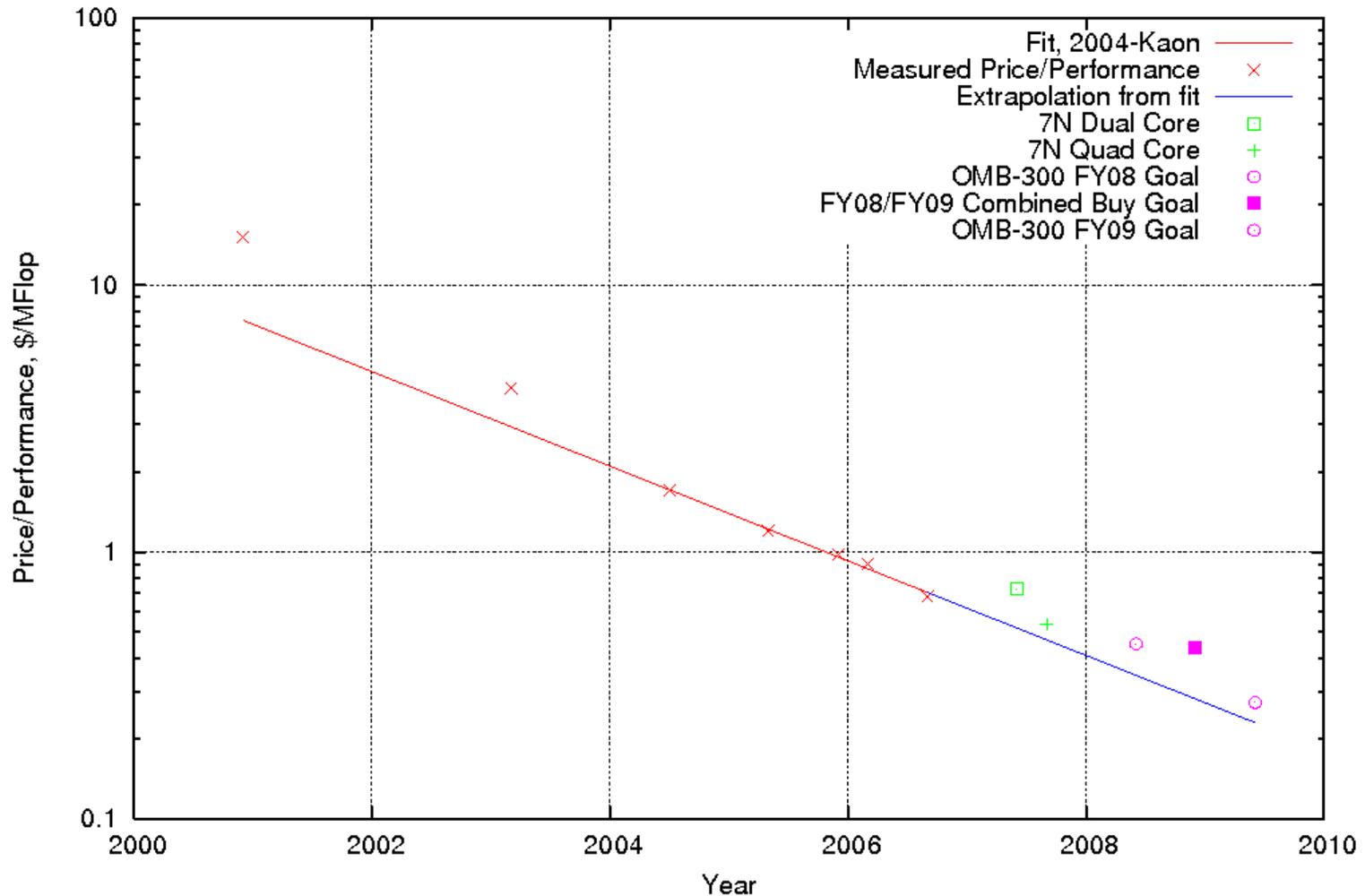
Performance on LQCD Codes

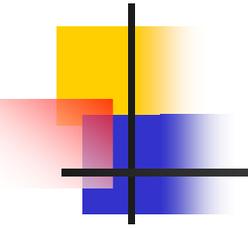
Performance of MILC Improved Staggered Inverter



Price/Performance Trends

Price/Performance of SciDAC/USQCD Clusters on Improved Staggered

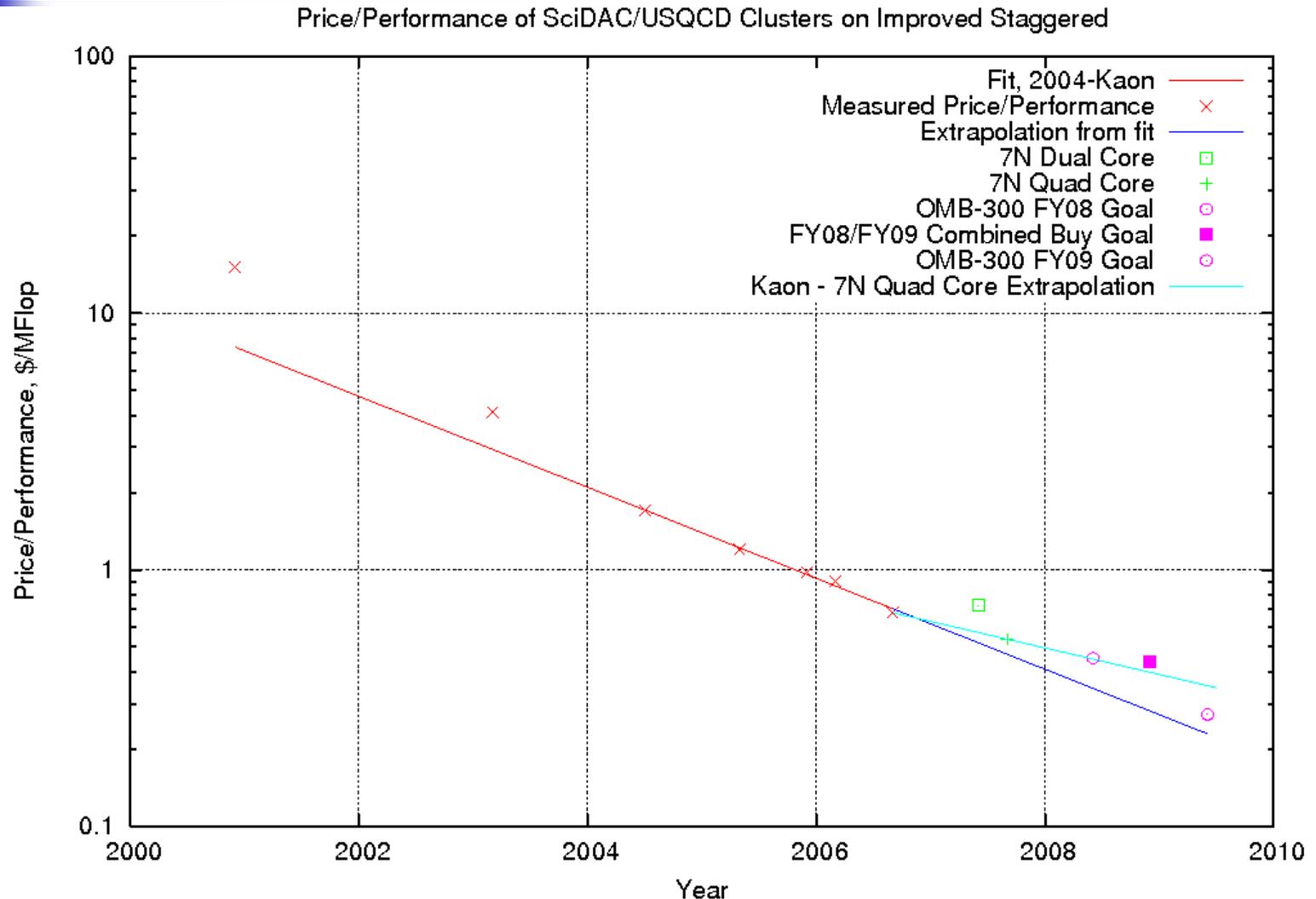




Extrapolating FY08/FY09 Performance

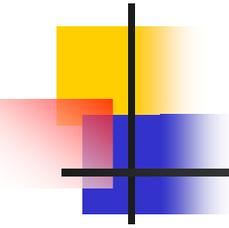
- Since 2004, thru the Kaon cluster, we have observed a halving time on price/performance on Asqtad on Myrinet and Infiniband clusters of 20 months
- 7N in both forms (dual and quad core) misses this trend line
 - Did we just have bad timing and miss a step function?
The Intel "Penryn" family that ships in August may be on the other side of the step.
 - Or, was there a slip for LQCD of 8-12 months in the downward trend that will now return to the trend slope?
 - Or, is this the start of a period with a longer halving time?
- Both new upcoming Intel processor/chipset generations directly address the memory bandwidth bottleneck that has affected LQCD performance (and the performance of many other major supercomputing application codes)
- We will have a much clearer picture by late Summer

Price/Performance Trend



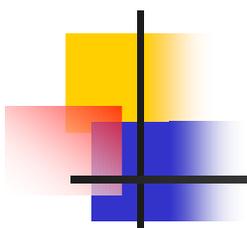
Impact on OMB-300 Delivered Tflops-yrs Milestones

- Currently deployed Tflops (BNL + FNAL + JLab):
 $4.2 + 3.6 + 1.2 = 9.0$
- Delivered Tflops-yrs milestones (FY06-FY09):
6.2, 9.0, 12.0, 15.0
- A conservative, achievable J-Psi estimate is 6.2 TFlops
- Achieved (**Predicted**) Tflops-yrs:
 - FY06: 6.21
 - FY07: 9.35 (extrapolating from data thru April, assumes no 7N)
 - FY08: 12.4 (assume 11.9 Tflops deployed, same pace as FY07)
 - FY09: 15.9 (assume J-Psi = 6.2 Tflops, JLab decommissioned 3G/4G, J-Psi operates for 9 months, only 90% uptime)
- Trend-based projected J-Psi performance range: 7.0 Tflops – 8.5 Tflops
(\$2.09M, based on measured trend, at FY08 and FY08/FY09 dates)



Summary

- We propose to combine FY08 and FY09 clusters into a single cluster, “J/Psi”, purchased via one subcontract
 - This increases delivered science in the first three years of operation and decreases procurement and integration labor
 - Deliveries in October and December 2008
 - Current funding profile is preserved
 - Release to production Jan 1 2009
- With a conservative prediction of 6.2 Tflops capacity for “J/Psi”, the project would exceed FY09 delivered Tflops goal of 15 Tflops
- Industry roadmaps indicate new architectures favorable to LQCD codes; if realized, achieved “J/Psi” performance will be significantly higher
- We will have solid performance measurements on the next generation of hardware in time to inform the FY09 OMB-300 submission

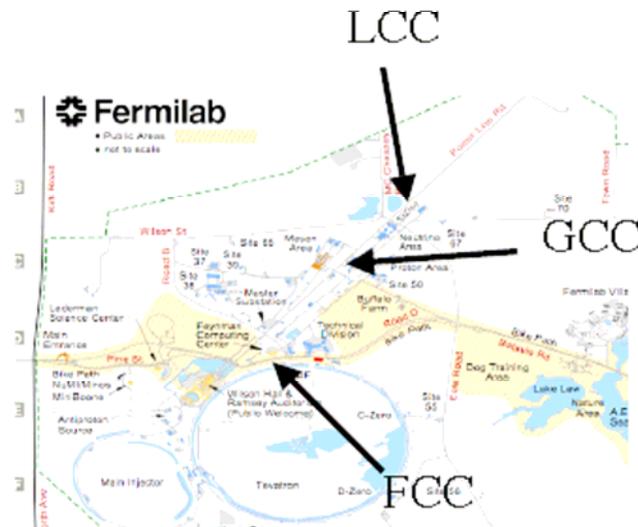


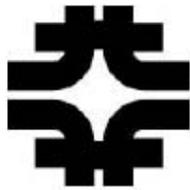
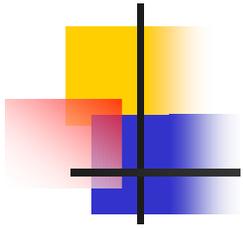
Backup Slides



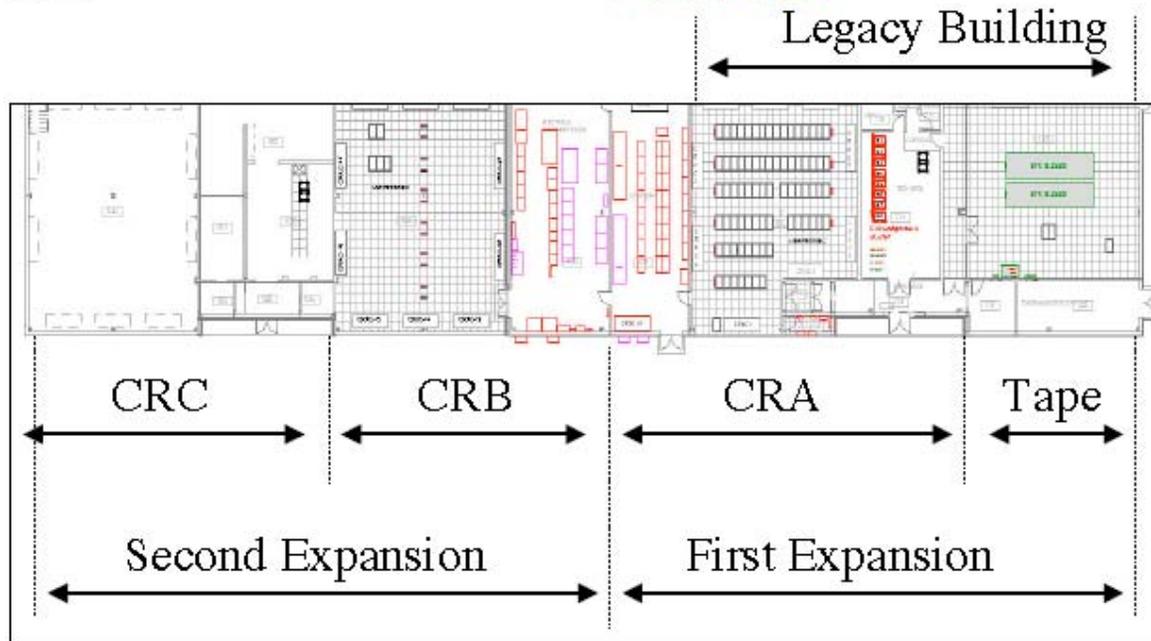
Computing Centers

- Three computing buildings.
- FCC, -- > 20 year old purpose built
- LCC, GCC: built on former experimental halls w/ substantial power infrastructure.





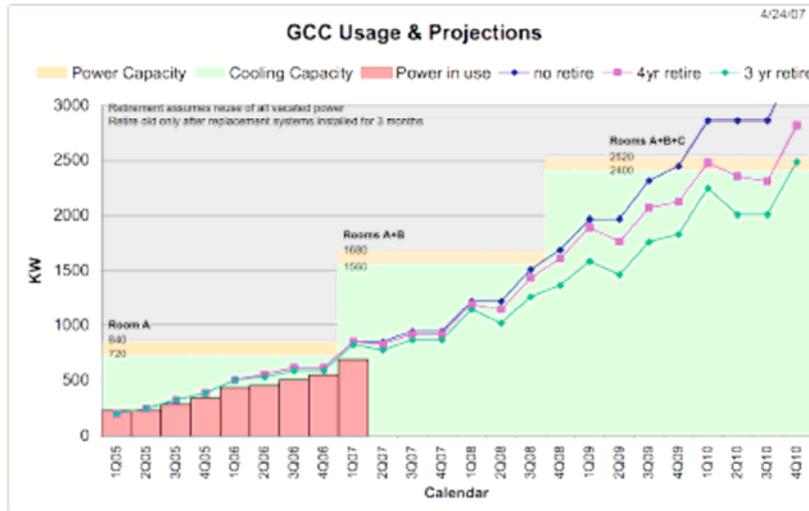
GCC Grid Computing Center



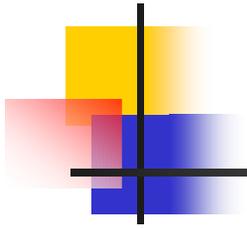


Just in Time Delivery

KVA, excluding cooling



FCC	600
LCC	710
CRA	840
CRB	840
CRC	840
Tape	45
Total	3875



Fermilab: Current Clusters

<u>Name</u>	<u>CPU</u>	<u>Nodes</u>	<u>Cores</u>	<u>Network</u>	<u>DWF</u>	<u>Asqtad</u>	<u>Online</u>
QCD	Single 2.8 GHz Pentium 4	127	127	Myrinet 2000	1400 MFlops per Node	1017 MFlops per Node	June 2004 0.15 TFlops
Pion	Single 3.2 GHz Pentium 640	518	518	Infiniband Single Data Rate	1729 MFlops per Node	1594 MFlops per Node	June 2005 / Dec 2005 0.86 TFlops
Kaon	Dual 2.0 GHz Opteron 240	600	2400	Infiniband Double Data Rate	4703 MFlops per Node	3832 MFlops per Node	Oct 2006 2.56 TFlops