LQCD Computing Project

Chip Watson, Jefferson Lab

Paul Mackenzie, Fermilab

acknowledgements to: Don Holmgren, Fermilab

3rd International Lattice Field Theory Network Workshop

Outline

- High Level Scope of the Project
- Near-Term Capacity Increases

  (Details: Paul Mackenzie, FNAL)
- 5 Year View

Project Scope

High Level View:

- Deployment of new resources
- Operations of new and old resources

Project Scope (1)

- Deployment
  - Annual installation of increasing computing capacity, matching the needs of increasingly challenging science problems
  - Aggregate capacity after 4 years of this project:
    ~18+ TFlops (double precision, sustained, average of key algorithms)
    Note that this is a conservative number to which we are willing to be held accountable by OMB.
  - Each year, we will select the best platform for the planned physics
    - Clusters are almost certain to be optimal in 2006 and 2007
    - Clusters are most likely to be optimal throughout the next 4 years, but we will track alternatives

Project Scope (2)

- Operations
  - Operate at three sites: BNL, FNAL, JLab
  - Operate as a metafacility
  - Operate multiple generations of new machines (an aggregate of 10-20 TFlops new in this project)
  - Operate the existing dedicated USQCD machines (~6 TFlops) as part of this metafacility

Current Capacity (1)

- QCDOC at BNL (Bob's talk)
  - 12K nodes with total of 4.2 teraflops
  - partitions sizes in teraflops:
    1.4, 0.7, 0.7, 0.35, 0.35, 0.15, 0.15

Current Capacity (2)

- Clusters at JLab
  - 256 node gigE – 0.2 teraflops
  - 384 node gigE – 0.5 teraflops, 3 partitions of ~0.15

  These will be collapsed into a single queue of 5 partitions of 128 nodes (more convenient to users)

Current Capacity (3)

- Clusters at Fermilab
  - 128 node myrinet – 0.14 teraflops
  - 260 node infiniband – 0.4 teraflops

Off Project Near Term Increases

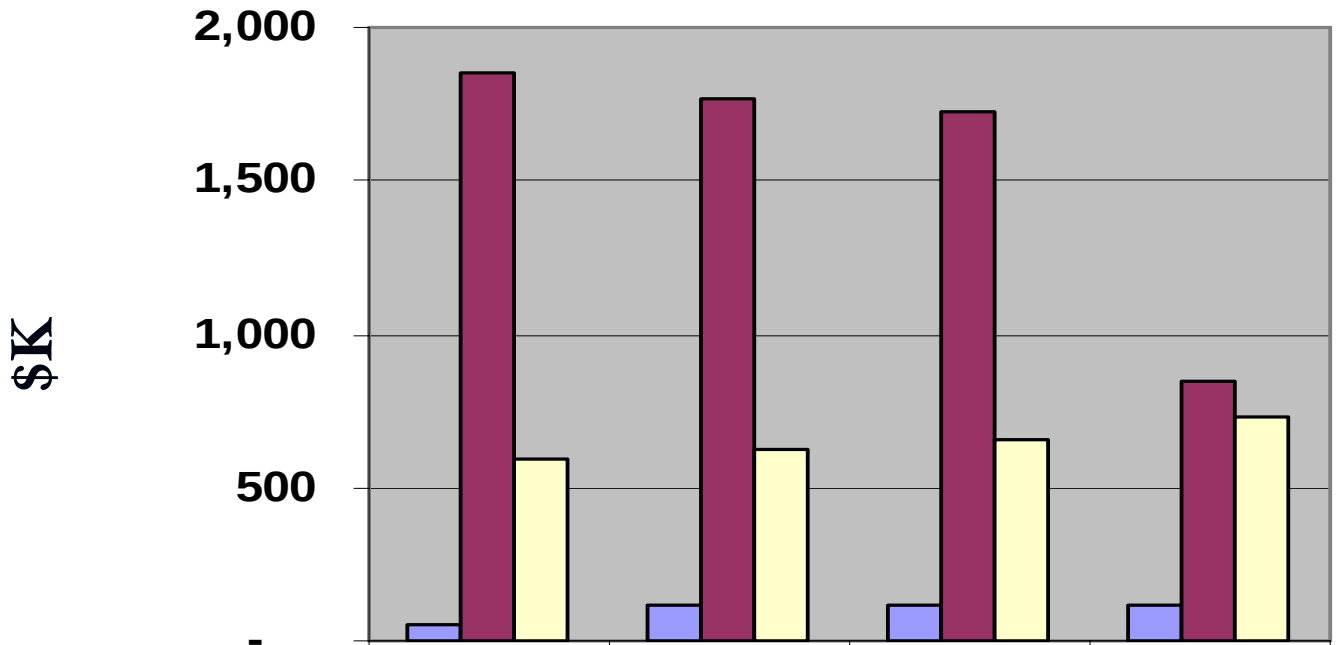
- Jefferson Lab
  - 130 node infiniband cluster by the end of 2005 (SciDAC + base funds)
  - this will double to 260 nodes (= 0.45 teraflops) once project funds are available
  - will evaluate Pentium D as alternative processor prior to procurement (alternative to existing FNAL nodes)
- Fermilab
  - 260 node infiniband cluster will double to 520 nodes by the end of 2005 = 0.9 teraflops (HEP funds)

(details: Paul's talk)

Budget Breakdown

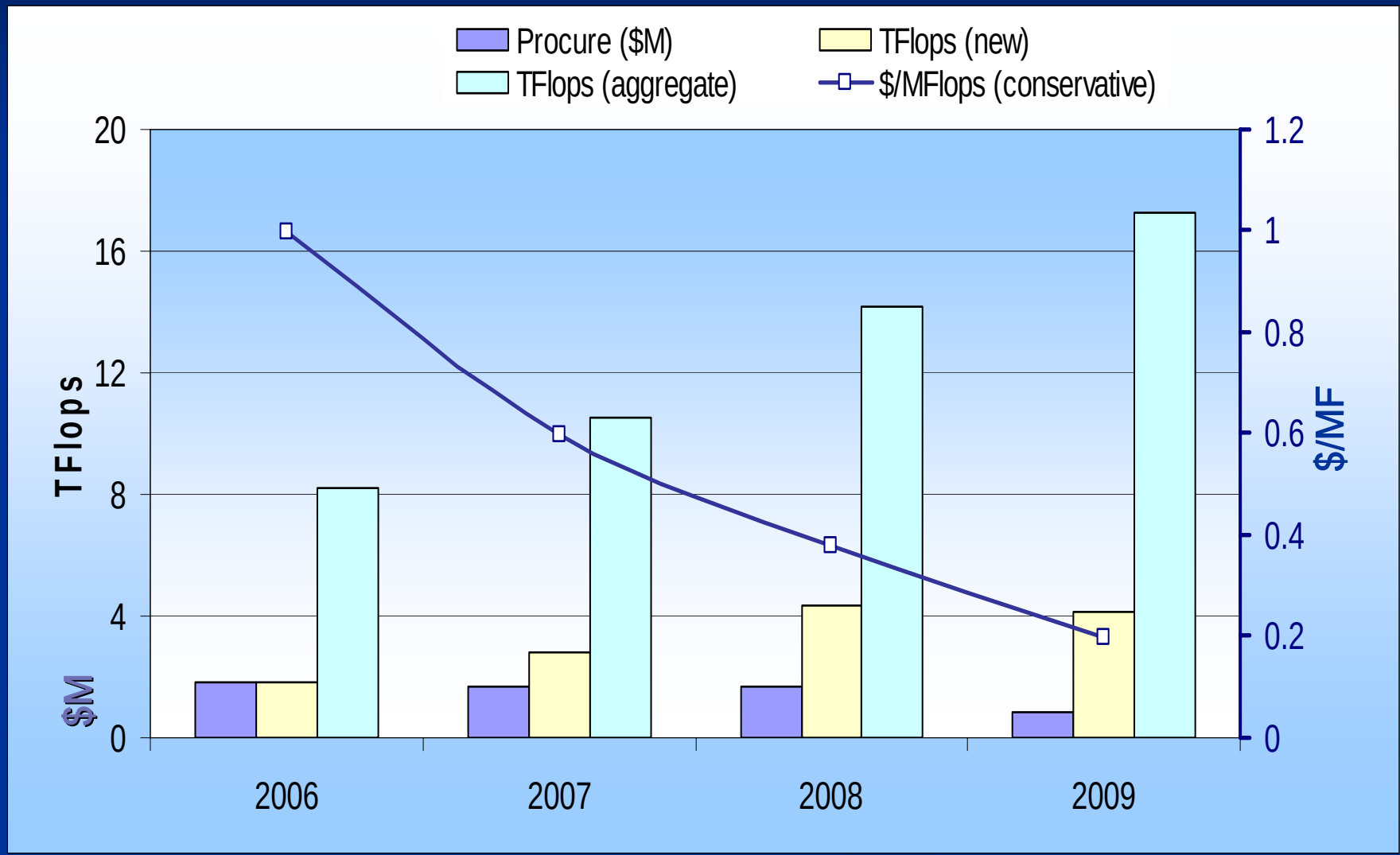
Each year, most of the funds will be used to deploy one large machine.

In the last year only a smaller machine is funded.



	FY06	FY07	FY08	FY09
Planning - DOE	55	114	119	119
Acquisition - DOE	1,856	1,766	1,730	852
Operation - DOE	589	620	651	730

LQCD Capacity Growth



Project Context

The following projects (and funding sources) significantly effect the new facilities project:

- SciDAC
- Base contributions (HEP, NP)
- ILDG – International Lattice Data Grid

Project Context : SciDAC

National Computational Infrastructure for Lattice Gauge Theory

Scope:

- Computing R&D (hardware platforms & software)
- API standardization, software implementation
- Performance optimization
- Prototype clusters: platform evaluations & software testbeds

Impact of SciDAC on this project:

- Lowers risk by reducing platform uncertainty
- Increases platform independence, code (& user) portability

Project Context (2)

- SciDAC-1 ends this year, but SciDAC-2 is assumed to continue for the life of this project (critical dependency)
- Ongoing base HEP and NP programs at the 3 labs
  - leveraged staffing and hardware (contributions)
  - leveraged computing infrastructure

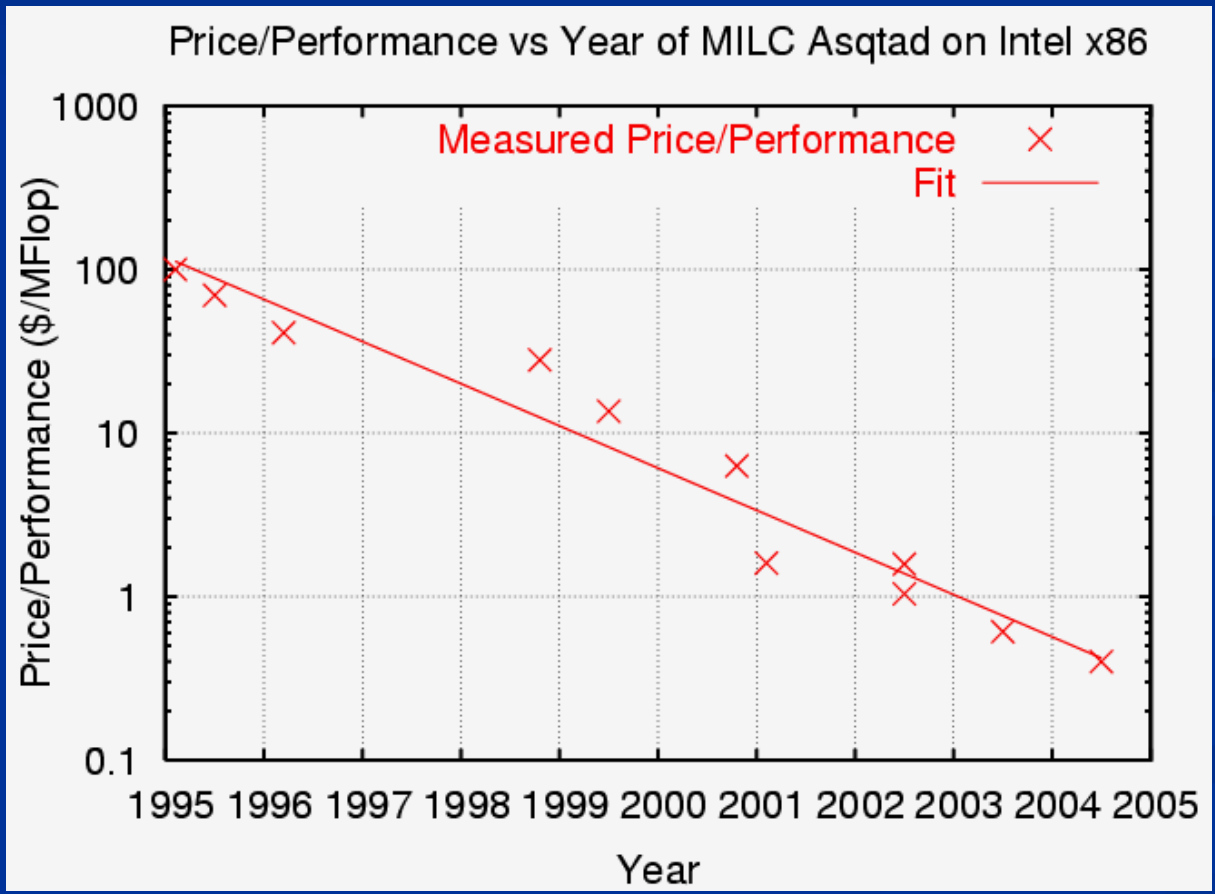
    (networking, tertiary storage, security, account management, ...)
- International Lattice Data Grid (ILDG)

  (significantly leveraged computational capacity; a data exchange)

2005-2006

(Paul Mackenzie)

Performance Trends – Single Node



MILC Improved Staggered Code ("Asqtad")

Processors used:

- Pentium Pro, 66 MHz FSB
- Pentium II, 100 MHz FSB
- Pentium III, 100/133 FSB
- P4, 400/533/800 FSB
- Xeon, 400 MHz FSB
- P4E, 800 MHz FSB

Performance range:

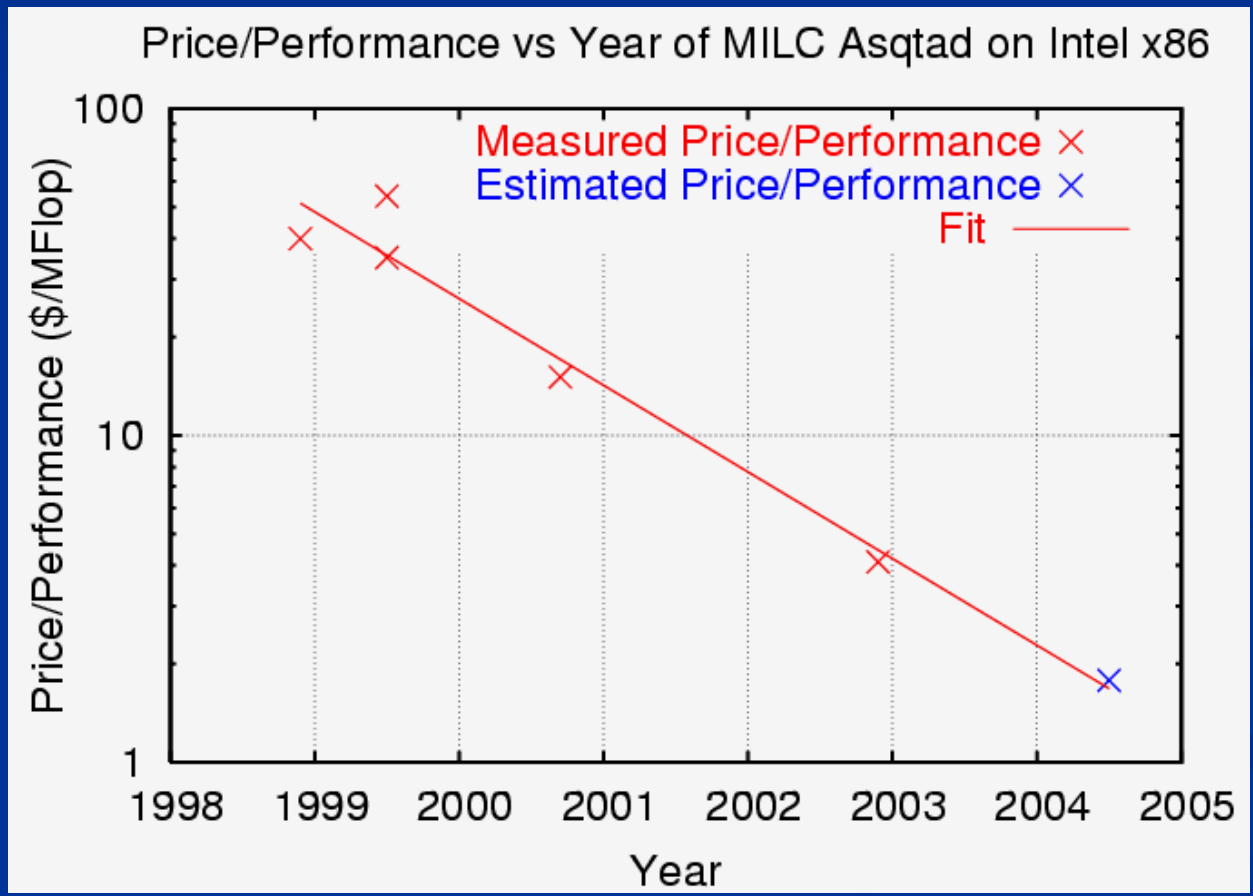
- 48 to 1600 MFlop/sec
- measured at 12^4

Doubling times:

- Performance: 1.88 years
- Price/Perf.: 1.19 years !!

Source: FNAL

Performance Trends - Clusters



Clusters based on:

- Pentium II, 100 MHz FSB
- Pentium III, 100 MHz FSB
- Xeon, 400 MHz FSB
- P4E (estimate), 800 FSB

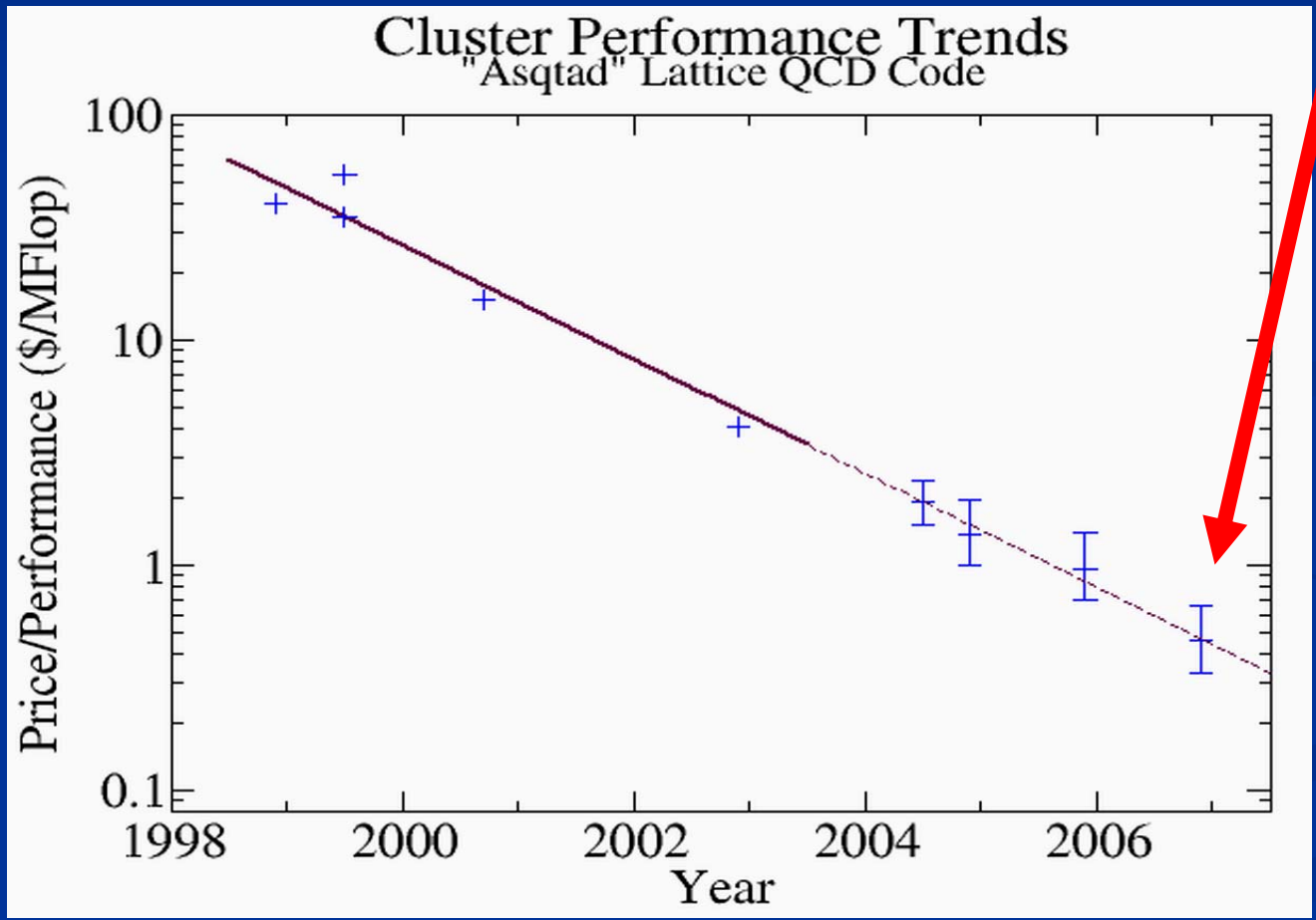
Performance range:

- 50 to 1200 MFlop/sec/node
- measured at 14^4 local lattice per node

Doubling Times:

- Performance: 1.22 years
- Price/Perf: 1.25 years

~1 Year Predictions



Mid to late 2006:

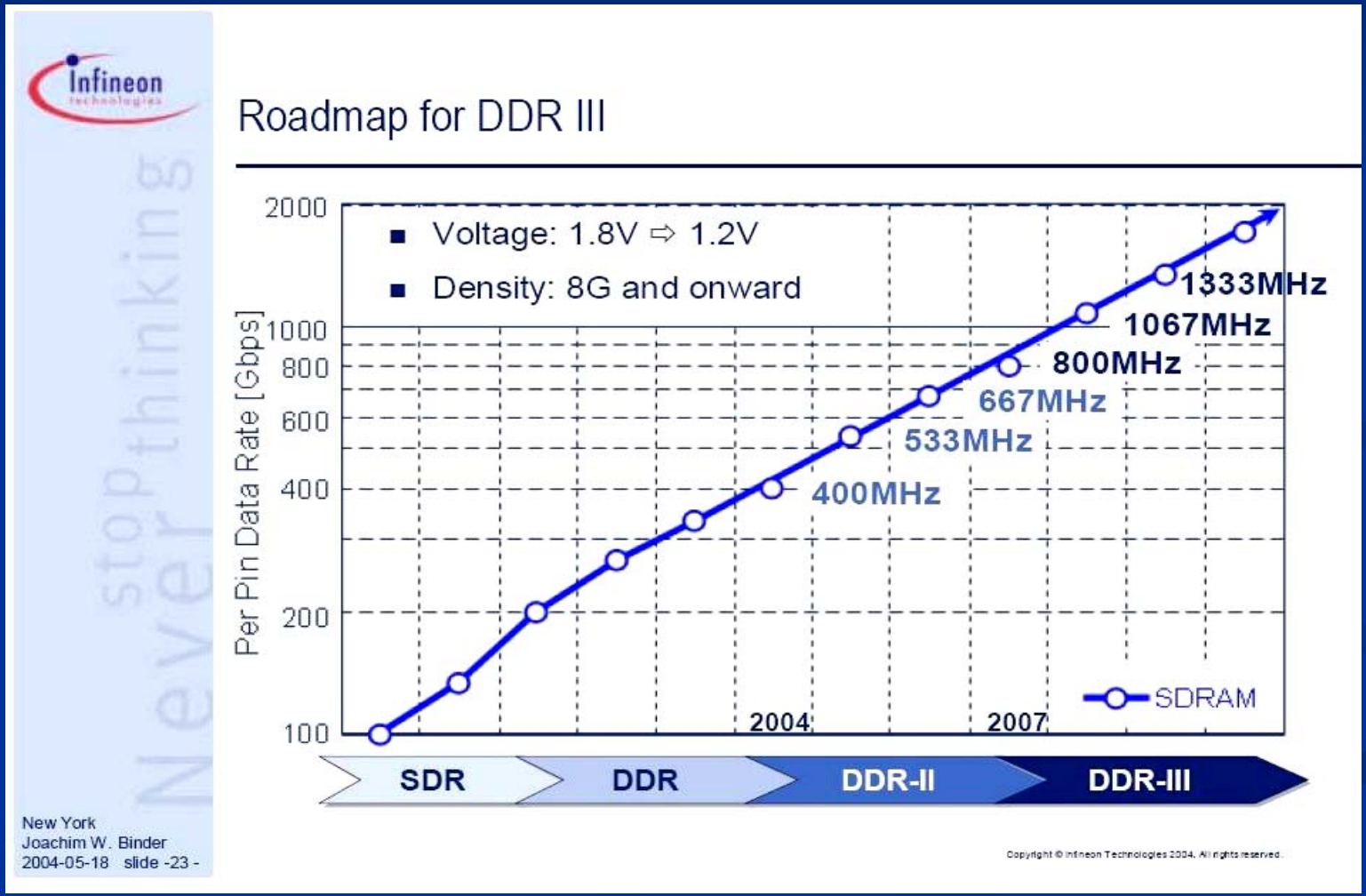
- dual core P4
- 1066 MHz FSB ("fully buffered DIMM technology")
- PCI-Express
- Infiniband
- $900 + $500 (system + network per node)
- 3.0 GFlop/node, based on faster CPU, higher memory bandwidth, cheaper network

Major future trends

Will these trends continue? Yes.

- Multi-core: SMP on-chip
  - 2005 / 2006: dual core, 2007: quad core
  - multi-core multiplies issue rate / clock cycle
  - but... exacerbates memory bandwidth issues
- Memory bus
  - going from 800 to 1066 now... to 1600 with a few years
  - going from shared bus to crossbar or NUMA (AMD)
- Cache size
  - 2MB today, 4MB in 2006 (2 per core), ...
  - as more of the problem fits into cache, more of the peak (potential) flops become sustained flops

Memory speed roadmap



High Speed Links (this decade)

- Infiniband
  - Infiniband 4x delivers 2 GBytes/sec bi-directional bandwidth at very low cost on PCI-Express
  - 4 to 6 usec latency (good enough for $2M - $3M machine)
  - network cost per node is falling rapidly (now < $800) and shows promise of falling considerably further
  - 8x (dual 4x) and 12x links exist; they will become mainstream as performance of boxes grows
  - end of decade: 12x DDR, 12x QDR? (>100 GBytes/sec)
- 10 gig ethernet
  - will provide price/performance pressure on Infiniband
  - latencies will be higher, but good enough for smaller clusters

High Speed Links - 2

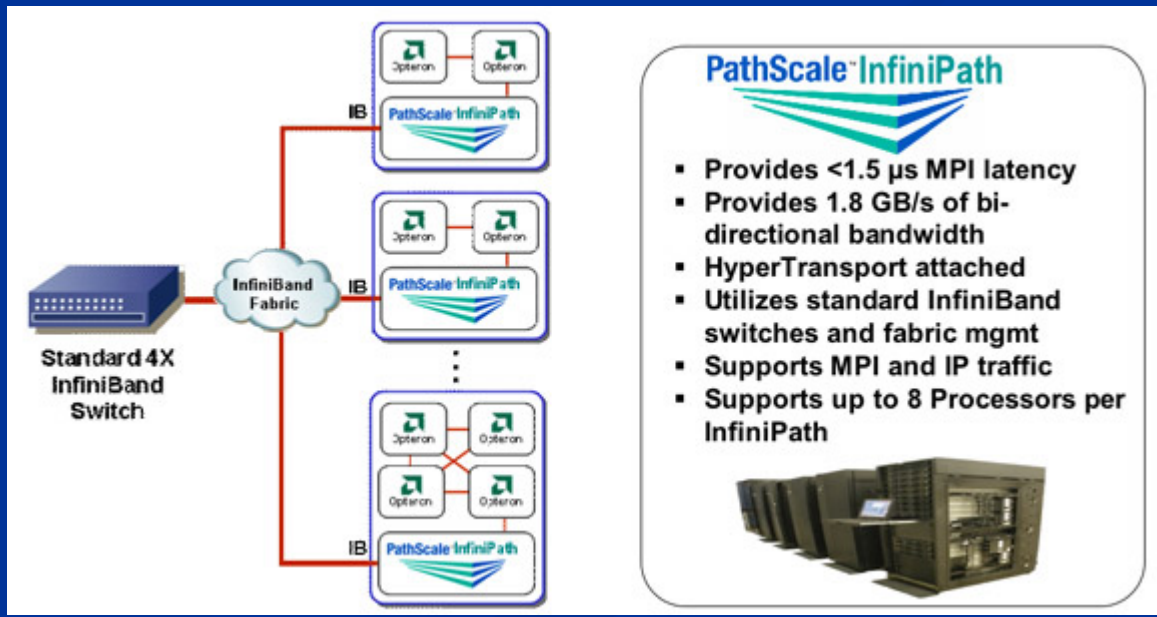
Pathscale Infinipath

- hypertransport to infiniband bridge
- 1.5 usec latency, 1+1 GBytes / second (growing)
- optimized for short messages ($n_{1/2}$ =600 bytes)
- direct from processor to I/O without going through memory!!!
- $70 / chip

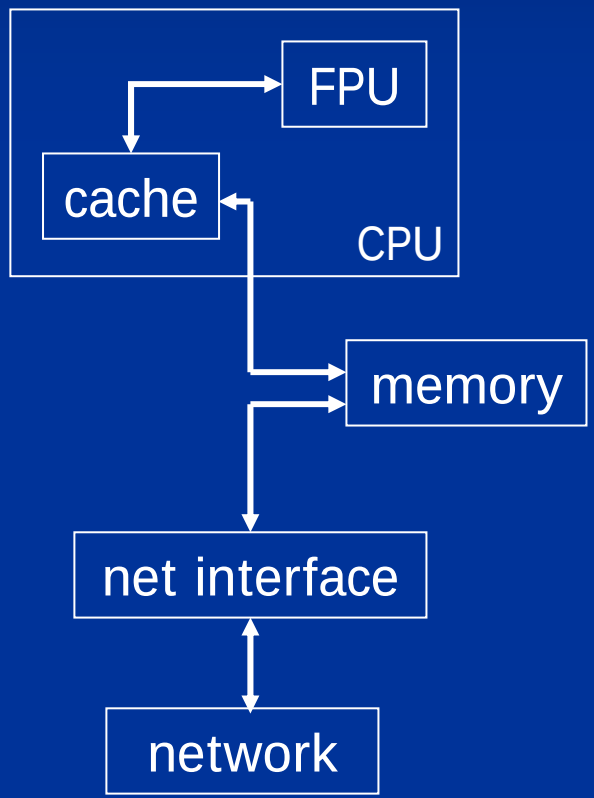
but...

- limited to AMD (today)

Hypertransport is a bus which can link CPUs and I/O devices, and is the native SMP bus for Opterons.

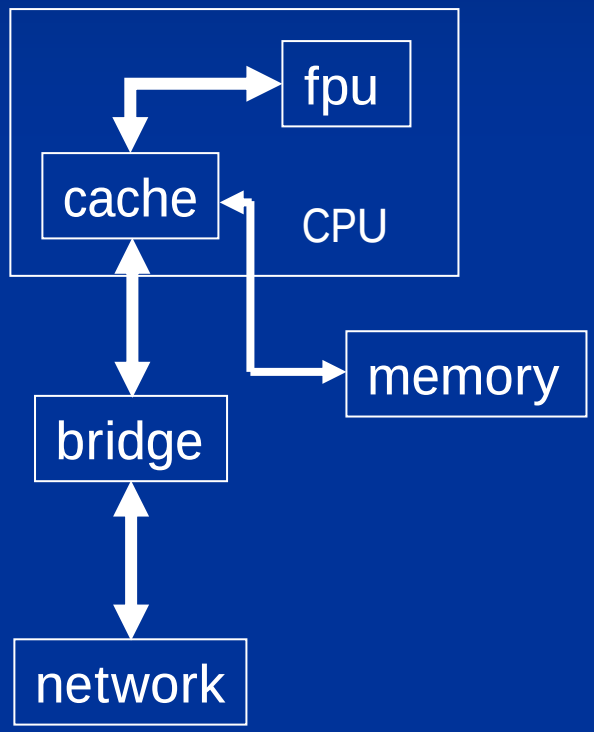


Classical memory bottleneck...



- Even for cache resident problem sizes, message data must cross the memory bus twice
- This limits network performance to ½ memory speed
- If message buffers must be built (scatter / gather), even more memory bandwidth is consumed in I/O

Getting around the bottleneck

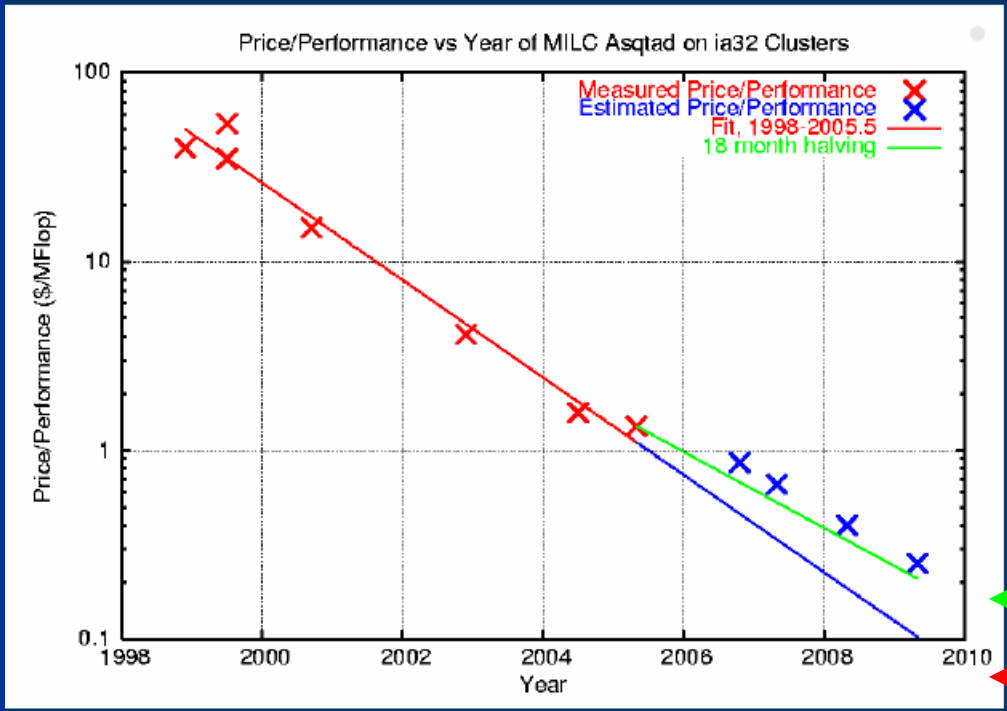


- the bridge chip sits in the processor's address space
- data can be written directly to the network, bypassing memory
- for multi-threading chips, one thread could do I/O
  (consumes one of the many threads available in the hardware; don't need to use a separate DMA engine)
- bandwidth limit is now no longer limited to memory speed, and I/O need not consume memory bandwidth

- Note: this is similar to KEK's work on the BlueGeneL torus!

4 Year Extrapolations

Performance Milestones - FY06-FY09



Measured and estimated asqtad price/performance

- Blue crosses derive from our "deploy" milestones
- Green line is Moore's Law with 18 month doubling time

Conservative

Trend line

Revolutions

While you can't schedule them, they do happen

- SSE: commodity vector processing
- Multi-core CPUs
- Pathscale (direct bridge from processor to network)
- Blue Gene L – "commodity" supercomputer?

5 Year Horizon

- Commodity
  - Infiniband (or replacement) network
  - Multi-core, likely SMP with multiple memory buses
    - Single processor possible if future network costs are similar to today's gigE
    - BlueGeneL successor – a 3D or 4D torus – also a possibility
  - Cluster trend line predicts $0.03 / MFlops in 2010
- Custom
  - Higher risk, aim at higher return
  - Must aim to beat commodity by 10x, so that schedule slips don't erase the gains (2 year slip = 3x in performance)
- Goal
  - 100 TeraFlops Sustained in 2010