

Teraflops and Beyond

Robert F. Lucas



High Performance Computing Research Department Head
National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
rflucas@lbl.gov

April 11, 2000

What is NERSC?

- The US Department of Energy, Office of Science, supercomputing center for the past 25 years
- Currently located at Lawrence Berkeley National Laboratory in the hills above the University of California, Berkeley campus





Outline

- A Teraflop today
- Ten Teraflops tomorrow!
- A Petaflop someday?
- Conclusions

Sixth International LS-DYNA Users Conference

3



Path to a Tflop/s at NERSC Cray C90 Installed In 1991

- Cray C90 installed in December 1991
- Stable high end production platform for seven years until 12/31/98



Sixth International LS-DYNA Users Conference

4

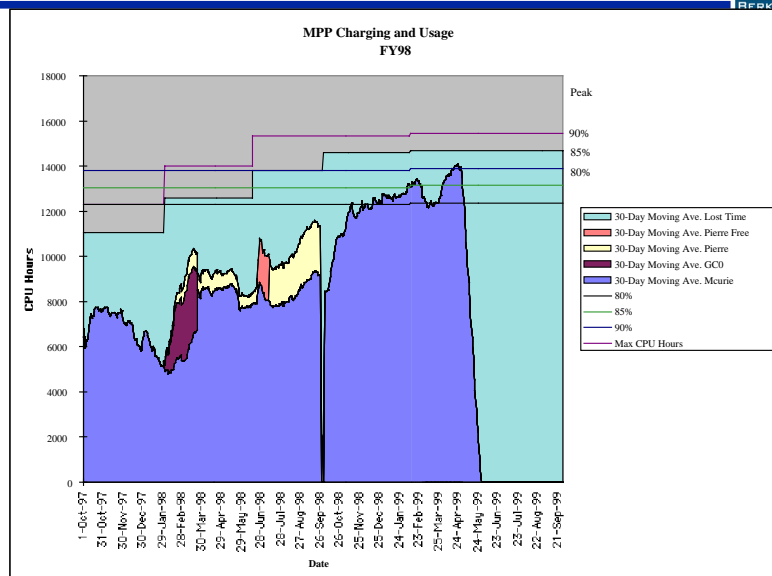
Getting Closer to a Tflop/s Cray T3E-900 Installed In 1996

The 696 processor T3E-900 is one of the most powerful unclassified supercomputers in the U.S.

1997 GAO report judged NERSC to have the best MPP utilization (75% then -- >90% today)



T3E Utilization





NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

We're Almost to a Tflop/s IBM SP3 Installed In 1999




BERKELEY LAB

- New contract with IBM announced in April 1999
- Phase 1 system delivered in June, 1999
- Phase 2 will exceed 3 Tflop/s in Q4 2000




Sixth International LS-DYNA Users Conference
7



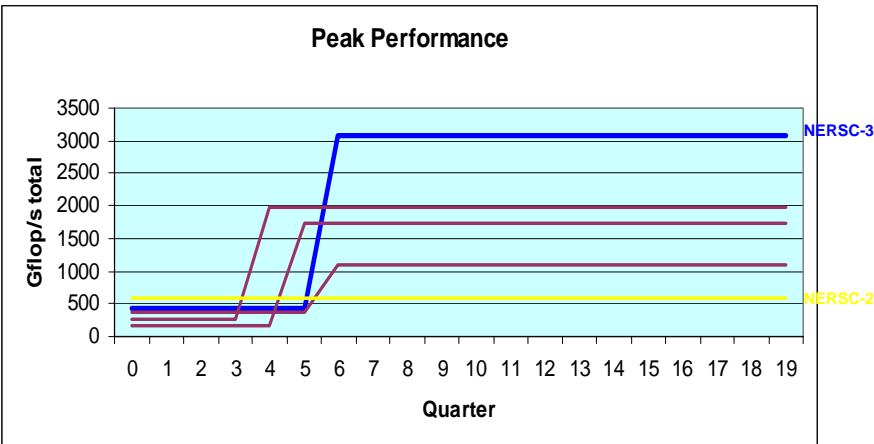
NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Why IBM?




BERKELEY LAB

Peak Performance




Quarter	NERSC-2 (Gflop/s total)	NERSC-3 (Gflop/s total)
0	~500	~500
1	~500	~500
2	~500	~500
3	~500	~500
4	~500	~2000
5	~500	~1800
6	~500	~3100
7	~500	~3100
8	~500	~3100
9	~500	~3100
10	~500	~3100
11	~500	~3100
12	~500	~3100
13	~500	~3100
14	~500	~3100
15	~500	~3100
16	~500	~3100
17	~500	~3100
18	~500	~3100
19	~500	~3100

Sixth International LS-DYNA Users Conference
8



ERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

What About Applications?




BERKELEY LAB

<u>Discipline</u>	<u>Computational technology</u>
magnetic fusion	particle in cell
computational chemistry material sciences	local density functional
climate research computational biology	partial diff. equations
QCD accelerator design particle detection simulation	Monte Carlo technique searching, pattern matching
combustion applied mathematics	PDEs image processing


Sixth International LS-DYNA Users Conference

9



ERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Good News: 1998 Gordon Bell Prize

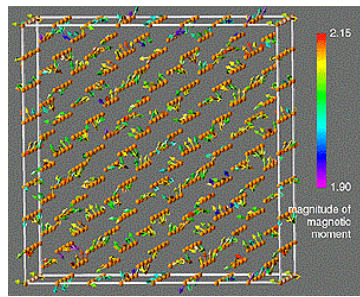


BERKELEY LAB

First complete application to break the 1Tflops barrier.



Collaborators from DOE's Grand Challenge on Materials, Methods, Microstructure, and Magnetism.

1024-atom first-principles simulation of metallic magnetism in iron



Sixth International LS-DYNA Users Conference

10





How About DYNA?

- World record is only 6 Gflop/s on a Fujitsu VPP 5000
 - VPP 5000 is a “cluster” of large vector nodes
 - Very high memory B/W per node
 - High B/W crossbar interconnect
- Lots of reasons why we’re not up to 1 Tflop yet:
 - Meshes are irregular and dynamically changing
 - Dot products and other mathematical bottlenecks to scalability
 - Note: complexity of LINPACK benchmark and LSMS Tflop/s application dominated by matrix-matrix multiply

Sixth International LS-DYNA Users Conference

11



Lets Look at Sparse Solvers

- Direct solver dominates the complexity of non-linear and dynamic analysis.
- Multifrontal algorithm is method of choice.
 - Dense arithmetic kernels -> maximize speed
 - Easily goes out-of-core -> maximize size
- Why is BCSLIB-Ext world record only 8 Gflop/s?

Sixth International LS-DYNA Users Conference

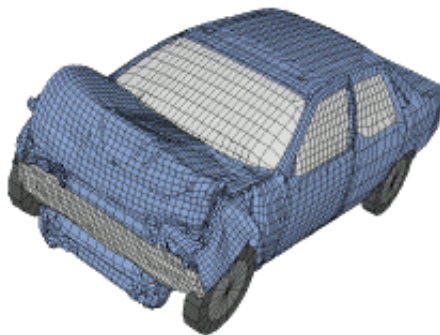
12

Load Balance



- Can't simultaneously load balance:
Finite Element Assembly
Factorization
Triangular Solves

Graph Partitioning

- How do you cut this grid in half?



- Optimal graph partitioning is NP-complete





Out-of-Core?

- What if you go out-of-core?
10,000,000 equation NASTRAN problems today!
- What system today provides 1 GB/s of standard I/O?
- Triangular solves completely I/O bound.
2 flops/word for linear solve
12 flops/word for block-shifted Lanczos

Sixth International LS-DYNA Users Conference

15





Scalable Solvers?

- Research codes ... one or more of the following usually apply:

Don't exploit symmetry!
Don't dynamically pivot!
Don't go out-of-core!
Don't map to initial distribution of finite elements!
Don't report performance of triangular solver!
- Usually benchmarked on trivial, 3D grids!

Sixth International LS-DYNA Users Conference

16





More Bad News

- The HPC community left Electrical Engineers behind twenty years ago.
SPICE didn't vectorize effectively
- Mechanical Engineers are falling by the wayside now
Not good for HPC industry
NASTRAN was the "killer app." for Crays
- Who's pushing the envelope?
Bigger problems?
Automatic optimization?
New algorithms?

Sixth International LS-DYNA Users Conference

17




Outline

- A Teraflop today
- **Ten Teraflops tomorrow!**
- A Petaflop someday?
- Conclusions


Sixth International LS-DYNA Users Conference

18




NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

What Will a 10 Tflo/s System Look Like?




BERKELEY LAB

- The first ones are already on order
Lawrence Livermore National Laboratory in US
Earth System Simulator in Japan
- Systems are large clusters
SMP nodes in US
Vector nodes in Japan
- Programming model:
OpenMP and/or vectors to maximize node speed
MPI for global communication




Sixth International LS-DYNA Users Conference

19



NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

A New Business Model?



BERKELEY LAB

Transition from vertical to horizontal companies

	sales		
applications software with MPI			
Irix	AIX		Solaris
Origin	SP	SPP	HPC
MIPS	PowerPC	PA-RISC	SPARC

→

	mail order		retail		
applications software with MPI					
Microsoft "NT"				Linux	
SGI	IBM	Compaq	HP	Sun	
Intel				others	

SGI

IBM

HP

Sun

Sixth International LS-DYNA Users Conference

20





Its Already Happening




- Forecast Systems Lab procurement
- Prime contractor is High Performance Technologies Inc. (HPTi)
- Subcontractor is Compaq

Sixth International LS-DYNA Users Conference 21





Its Really Nothing New!

- Remember EDS?
- Integration is how Ross Perot got rich!



Sixth International LS-DYNA Users Conference 22





Buying From Integrators Worries Me

- Integrator's don't "own" the whole system end-to-end.
- Who takes responsibility for problems?
- Who fixes third party H/W or S/W bugs?
- Who adds new features not demanded by transaction, database, or Web server markets?
 - Checkpointing
 - Low-latency communication?

Sixth International LS-DYNA Users Conference

23

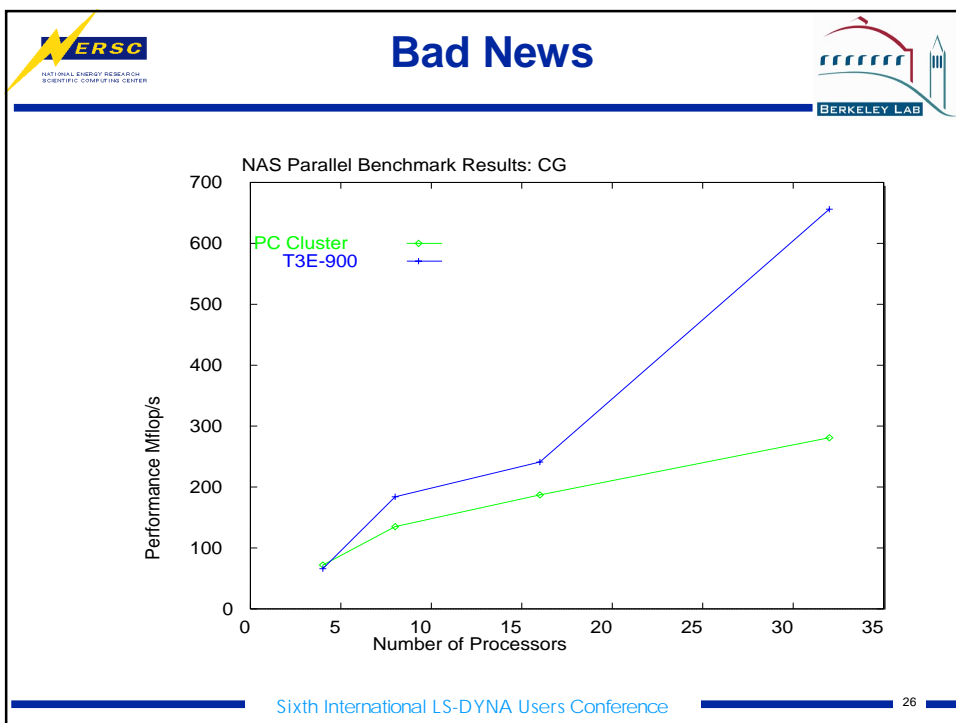
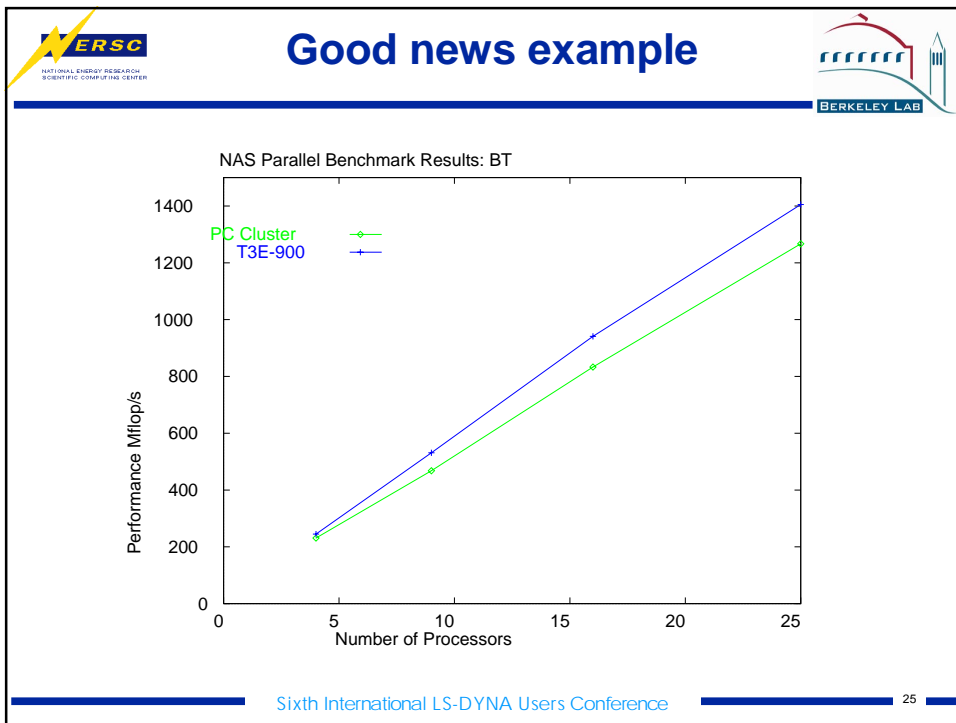


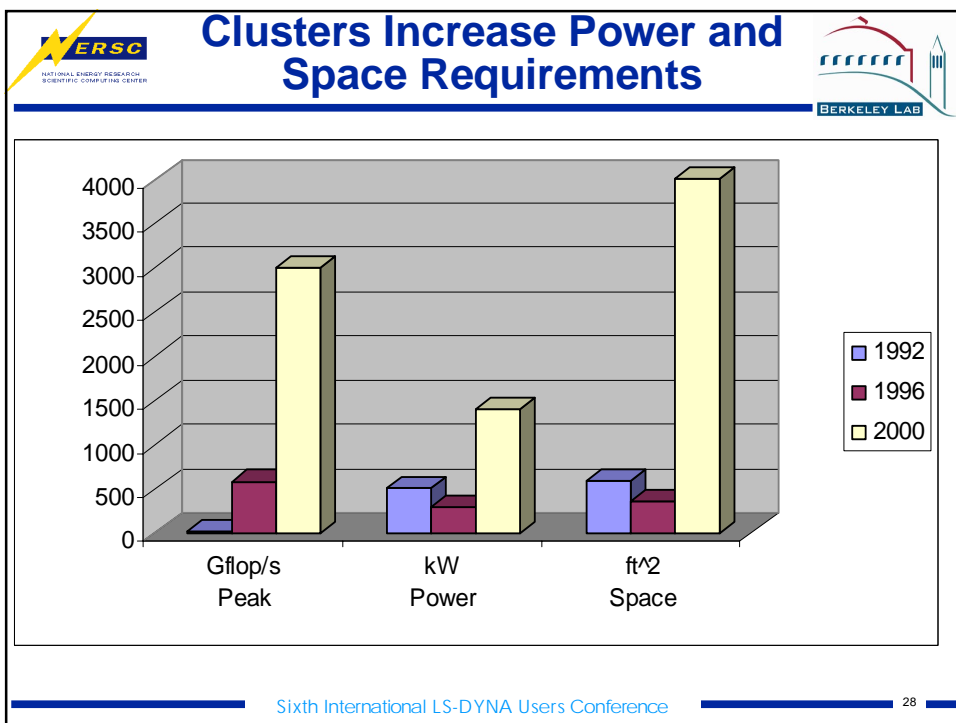
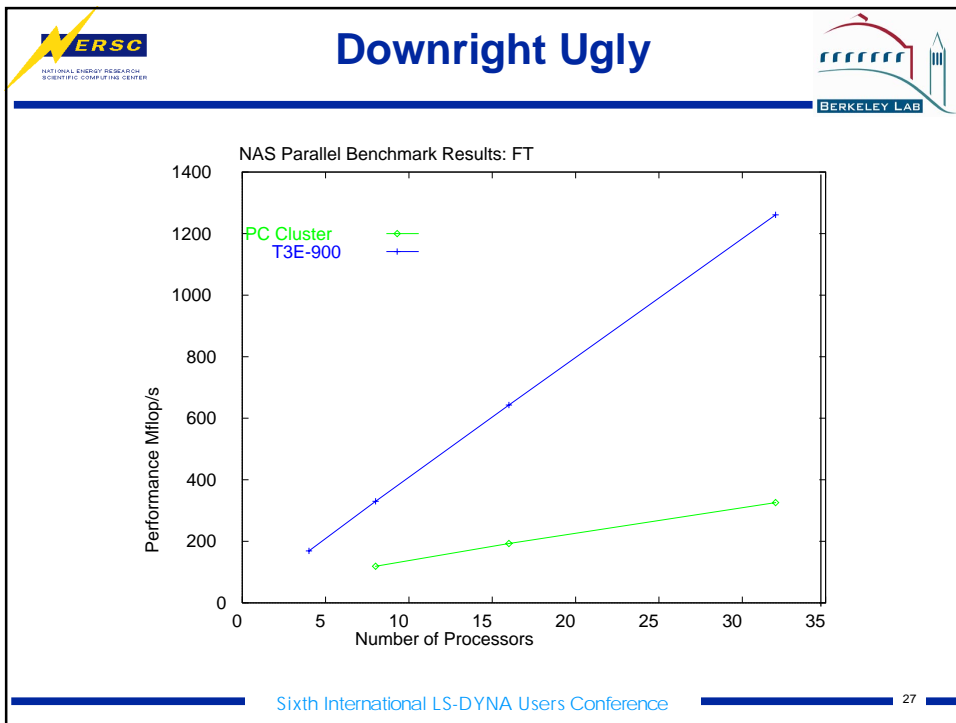
How Do Clusters of SMPs Perform?


- Good news:
 - Cheap flop/s and memory
 - PC's in LSTC's classroom faster than LSTC's 8-processor Origin 2000 on an explicit problem
- Bad news:
 - Global communication B/W is low
 - Communication latency is high
 - Net result will be difficulty with implicit solutions

Sixth International LS-DYNA Users Conference

24






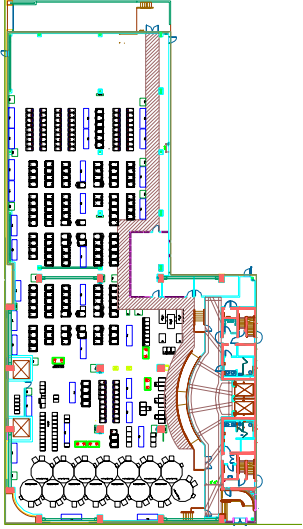


NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Eight Tflop/s Floorplan




BERKELEY LAB




Sixth International LS-DYNA Users Conference

29




NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

We're Not Quite There Yet





BERKELEY LAB



Sixth International LS-DYNA Users Conference

30





Near-term Prediction

- Peak performance and memory will go up
- So will power and floor space
- Most big systems will be procured from integrators
- The space of applications enjoying this will shrink
- The space of applications will certainly change
 - Materials!
 - Designer drugs?
- Implicit codes will probably not be well served at the very high end

Sixth International LS-DYNA Users Conference

31




Outline

- A Teraflop today
- Ten Teraflops tomorrow!
- A Petaflop someday?
- Conclusions


Sixth International LS-DYNA Users Conference

32



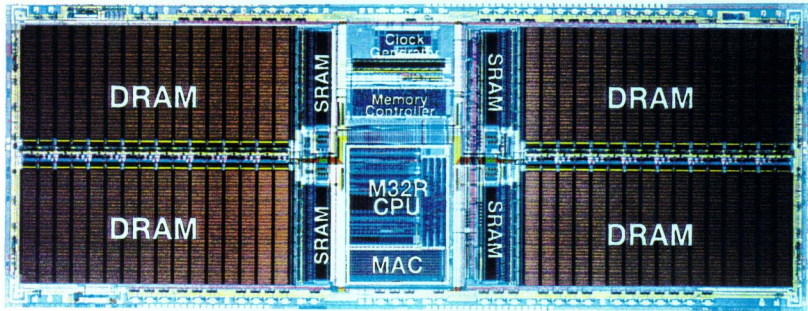
NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Processor in Memory Chips




BERKELEY LAB

- Processing in memory is an old idea
- Semiconductor technology now makes it very attractive




Mitsubishi M32R

Sixth International LS-DYNA Users Conference
33



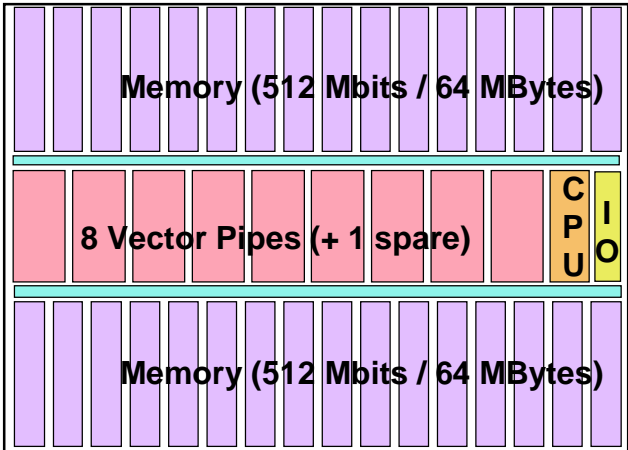
NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Rediscovering Vectors




BERKELEY LAB

- Couple this with vectors for high BW




UC Berkeley IRAM floorplan

Sixth International LS-DYNA Users Conference
34



ERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

CMOS Petaflop/s Solution




BERKELEY LAB

- 100,000 10 Gflop/s vector PIM chips
- 3D mesh interconnect
- $O(10^7)$ operations must be sustained on every clock cycle to avoid an Amdahl bottleneck
- IBM Blue Gene is an example
- God help the programmers!


Sixth International LS-DYNA Users Conference

35



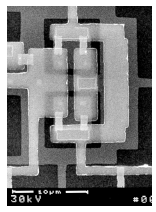
ERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

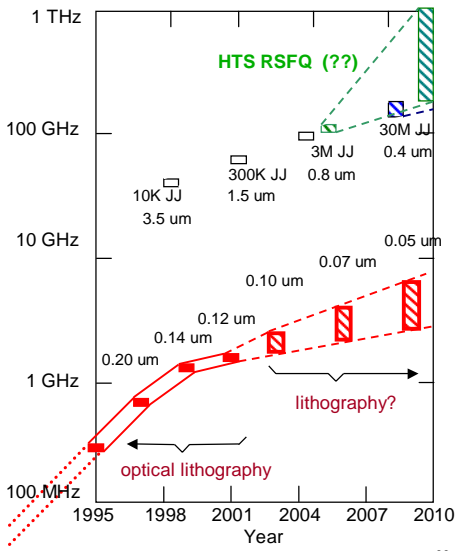
An Alternate Technology?



BERKELEY LAB

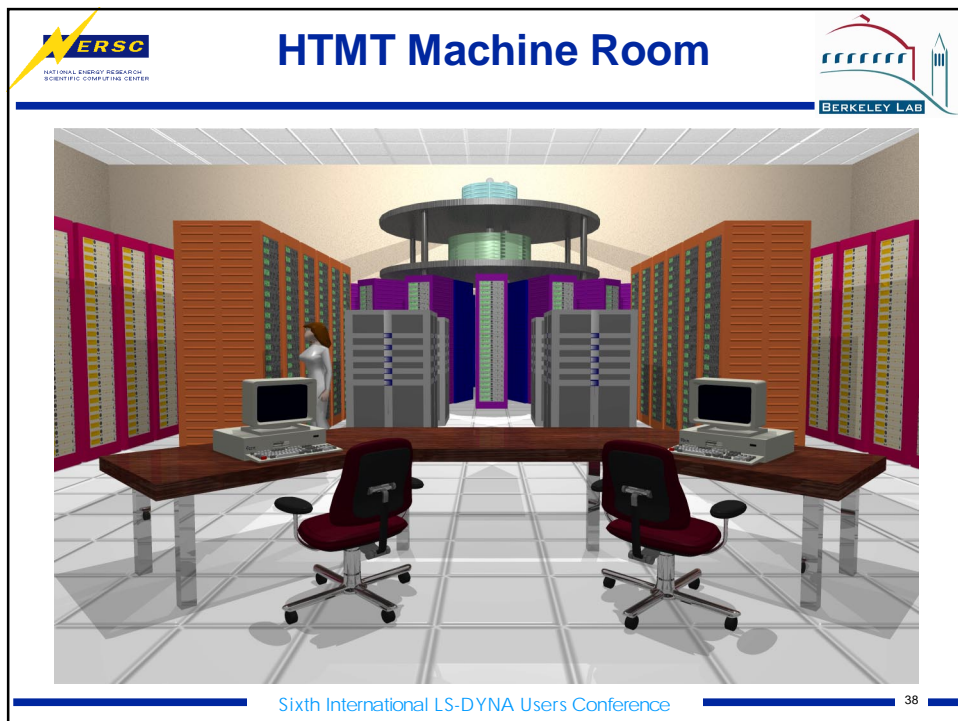
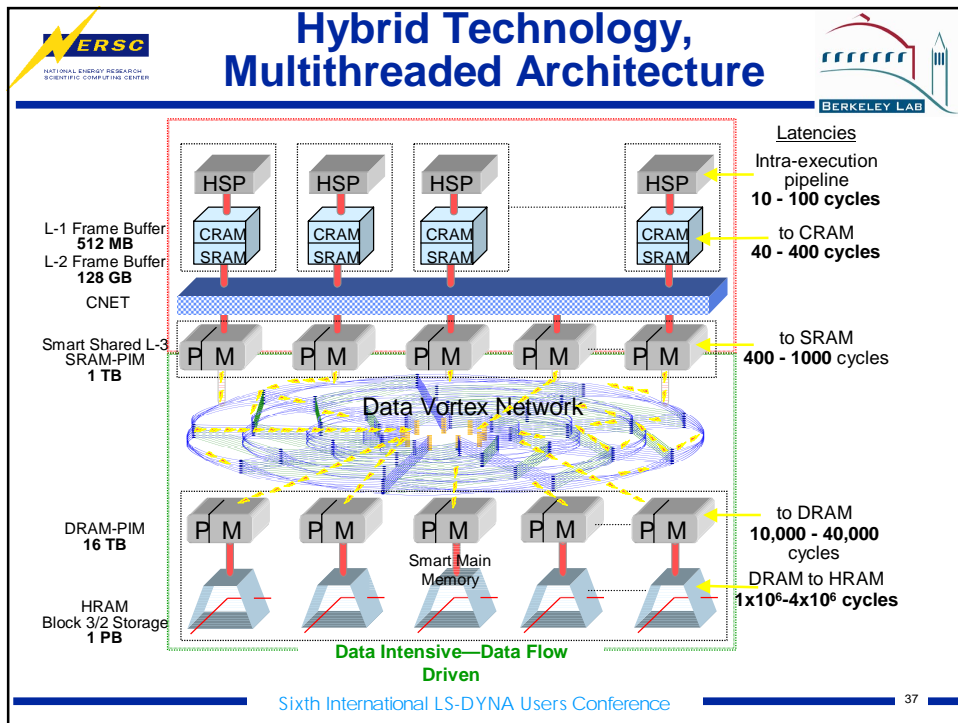
- Single Flux Quantum (SFQ)
- Operates at 4 Kelvin

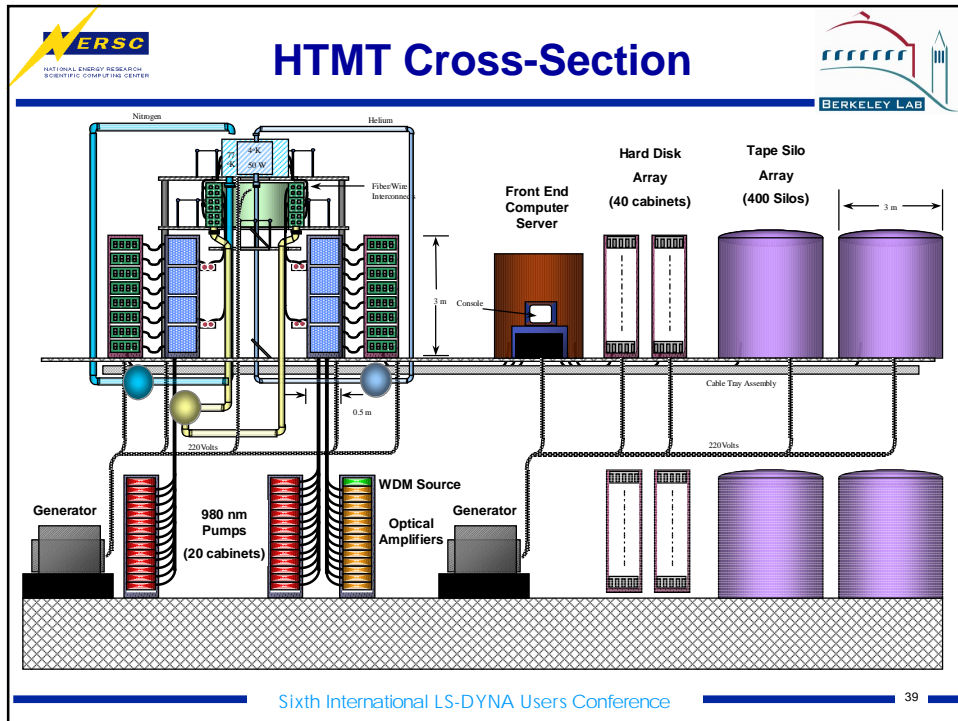




Sixth International LS-DYNA Users Conference

36





Outline

- A Teraflop today
- Ten Teraflops tomorrow!
- A Petaflop someday?
- **Conclusions**

ERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

BERKELEY LAB

Sixth International LS-DYNA Users Conference

40

- For most of us:
 - PCs will become SMPs
 - Larger, more complicated analysis will be common
- At the very high end:
 - Clusters of SMPs will predominate for the next five years
 - Peak performance will keep going up
 - It'll get harder and harder to realize
- Long Term Future:
 - Architecture will change again
 - Programming model will change again
 - Application base will change again

