BERKELEY LAB

Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY

UNIVERSITY OF CALIFORNIA

The Future of Large Scale Visual Data Analysis

Joint Facilities User Forum on Data Intensive Computing

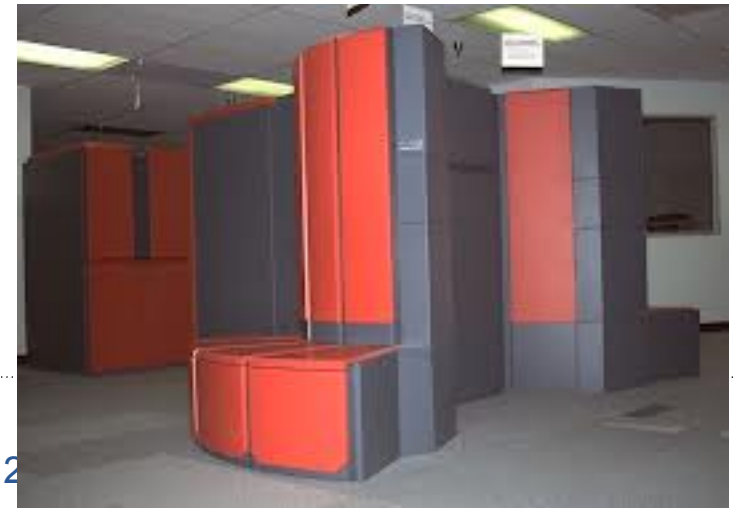
Oakland, CA

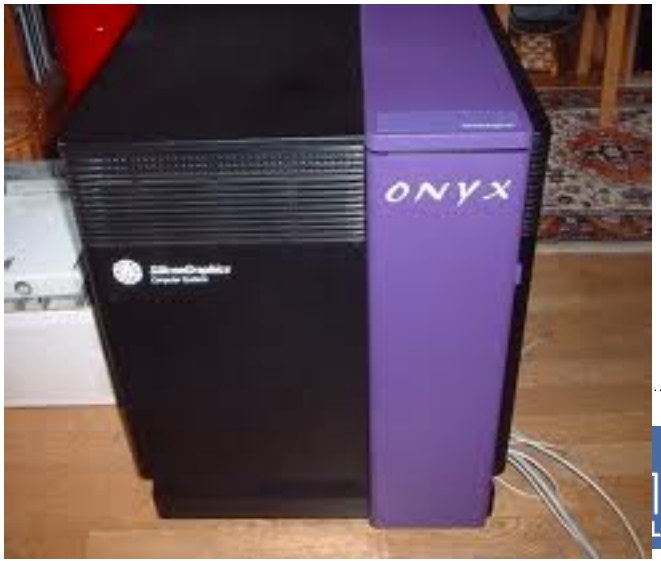
E. Wes Bethel
Lawrence Berkeley National Laboratory

16 June 2014

The World that Was: Computational Architectures

- Machine architectures
  - Single CPU, single core
  - Vector, then single-core MPPs
  - “Large” SMP platforms
  - Relatively well balanced: memory, FLOPS,I/O





The World that Was: Software Architecture

- Data Analysis and Visualization (DAV) Software
  - Subroutine-callable libraries
  - MPI-per core executables
  - And a generation of single-threaded apps

NCAR graphics
VisIt, ParaView
ferret, CDAT, gnuplot
AVS, DX, …

The World that Was: Use Models

- *Post Hoc*
  - Simulations save data to disk
    - Question: how much support to uses have centers given for parallel I/O over the years? (footnote)
  - Later, have a look at what was saved
  - Some noteworthy exceptions:
    - Cactus – PSE for building codes and plugging in "thorns" that do vis/analysis
    - CUMULVS (ca 2004) – computational steering/vis
    - Other custom solutions

The World that Will Be: Slide of Doom (1)

System Parameter	2011	"2018" Swim Lane 1	"2018" Swim Lane 2	Factor Change
System Peak	2 Pf/s	1 Ef/s		500
Power	6 MW	≤20 MW		3
System Memory	0.3 PB	32–64 PB		100–200
Total Concurrency	225K	1B×10	1B×100	40,000–400,000
Node Performance	125 GF	1 TF	10 TF	8–80
Node Concurrency	12	1,000	10,000	83–830
Network BW	1.5 GB/s	100 GB/s	1000 GB/s	66–660
System Size (nodes)	18700	1,000,000	100,000	50–500
I/O Capacity	15 PB	300–1000 PB		20–67
I/O BW	0.2 TB/s	20–60 TB/s		10–30

Aggregate concurrency grows by O(5-6)
Memory grows by O(2): less memory per core.
I/O capacity, BW grows by O(1): can't save all data.

The World that Will Be: Use Models

For computational and experimental science:

- *Post hoc.* There will always be data products.
- *In situ.* Do vis/analysis while data still resident in memory.
- *In transit.* Do vis/analysis on a “nearby machine”, but don’t save to storage first.
- *Workflow, work orchestration.* Sequences of compute and data-centric operations.

Implications of Changing Architecture

- Vis/analysis codes need to be retooled to operate on new architectures
  - Many more cores/processor
  - Much less memory/core than in the past
  - Power constraints
- Likely to be as “disruptive” as the phase-change from scalar to MPP
- Doing MPI per core won’t work, explicit threading unlikely to work.

The Cost of MPI per core

Howison, Bethel, Childs. MPI-hybrid parallelism for volume rendering on larrge, mult-core systems. EGPGV, 2010.

- Per PE memory:
  - About the same at 1728, over 2x at 216000.
- Aggregate memory use:
  - About 6x at 1728, about 12x at 216000.

Cores	Mode	MPI PEs	MPI Runtime Memory Usage		
			Per PE (MB)	Per Node (MB)	Aggregate (GB)
1728	MPI-hybrid	288	67	133	19
1728	MPI-only	1728	67	807	113
13824	MPI-hybrid	2304	67	134	151
13824	MPI-only	13824	71	857	965
46656	MPI-hybrid	7776	68	136	518
46656	MPI-only	46656	88	1055	4007
110592	MPI-hybrid	18432	73	146	1318
110592	MPI-only	110592	121	1453	13078
216000	MPI-hybrid	36000	82	165	2892
216000	MPI-only	216000	176	2106	37023

The Cost of MPI per core

- F
- A

Lessons learned:

- Doing MPI-per-core is not a sustainable solution at extreme scale
- MPI+X runs faster, uses less memory, moves less data.

Thought about the future:

- Likely the case that explicit threading will run into the same barriers: limits caused by the weight of the overhead.
- Implicit parallelism (e.g., data parallel) holds much promise (e.g., CUDA does this on GPUs)

Co						3)
17						19
17						13
138						51
138						65
466						18
466						07
1105						18
1105						78
216000	MPI-hybrid	36000	82	165	2892
216000	MPI-only	216000	176	2106	37023

Implications of Changing Use Models

- Doing full-resolution data saves for *post hoc* analysis/vis likely not practical (possible?)
- Migration from *post hoc* to *in situ*
  - Codes need to be retooled:
    - Past: calls to I/O library
    - Future: calls to *in situ* infrastructure (footnote)
  - Implications for sharing limited resources
    - Cores, memory, data movement, power budget

Overview of *In Situ* Infrastructure

ADIOS	Code modification required	I/O based, user-pluggable processing, can do I/O, runtime configurable, non-zero copy, inline data transformations, staging.
GLEAN	No code modification required	I/O intercept, user extensible analysis (via the GLEAN API), staging.
VisIt/Libsim	Code modification required	Tightly coupled, zero-copy (in progress), connects simulation to VisIt client.
ParaView/ Catalyst	Code modification required	Tightly coupled, zero-copy, connects simulation to ParaView.

Implications of Changing Use Models

- Increasing emphasis on complex workflows (productivity)
  - Coupling between simulation, experiment
  - End-to-end view of data solutions
    - Data management, processing, movement, analysis, vis, sharing/publishing, curation
  - Automation of formerly (presently?) manual operations

How is the community responding?

- Increasing portability and parallelism.
  - Several research projects focusing on DSL-like approach for expressing algorithms, achieving high concurrency and platform portability (DAX, EAVL, PISTON, etc)
    - Note: the same kind of thing is happening across many communities, including ML
- Infrastructure for legacy and future applications?
  - Problem: VisIt and ParaView in widespread use
  - Solution: SciDAC3 SDAV & vtk-m (2-3 yrs out)

How is the community responding?

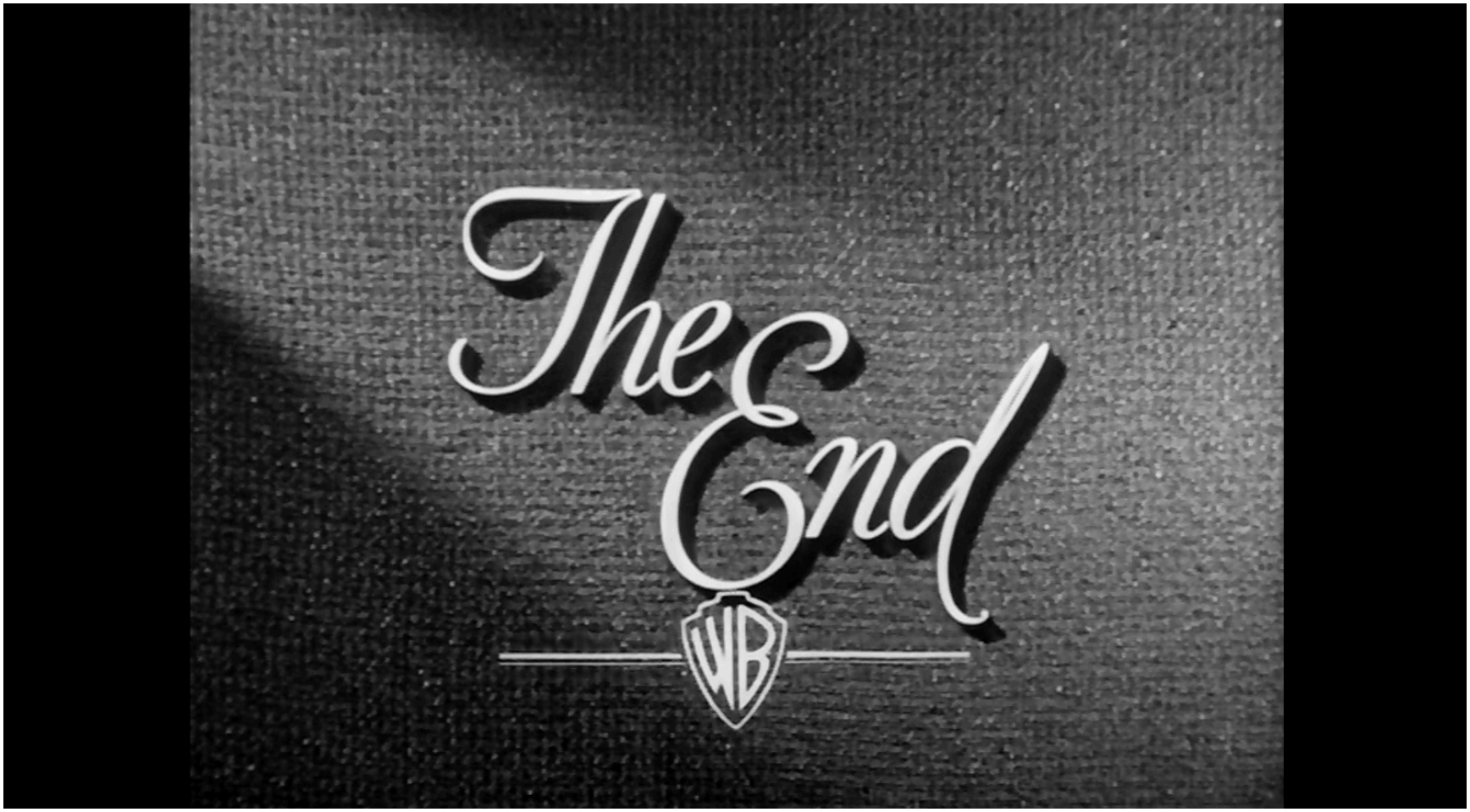
- 5-10 years out
  - *In situ* infrastructure matures
  - Less distinction between “analysis” and “vis”
    - It may be data features or statistics that are viewed rather than raw field/particle/mesh data
    - Analysis of flow (e.g.), want to “see” analysis results
  - Evolving data software stack
    - Accommodates major exascale challenges: resiliency, power, portability, resource mgt

Future of Large Scale Visual Data Analysis

- Code teams and *in situ:*
  - "Resistance is futile."
- Computing facilities:
  - Users will need help with *in situ,* workflow infrastructure.
    - Question: how support to users have centers provided over the years for parallel I/O?
  - The future data-centric software will be much more complex than what you've seen in the past.

Future of Large Scale Visual Data Analysis

- Vis/analysis infrastructure will be ready for future architectures
  - This very subject consumes a large fraction of R&D funding.
- Partnering with facilities and code teams is a key element of achieving that objective
  - Data-centric projects/pilots help push the limits of technology and prepare you for the future.


The End
WB