“Architecture Recapitulates Phylogeny”

How Scalability Requires Specialization

Steve Oberlin

10/9/2001

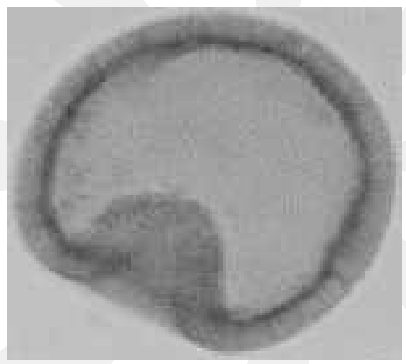
UNLIMITED SCALE, INC.

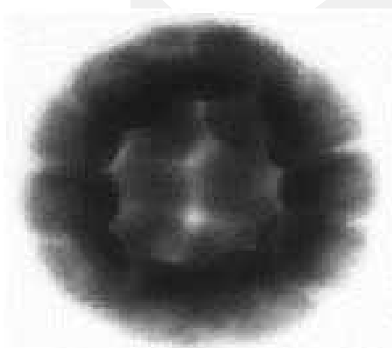
About the title...

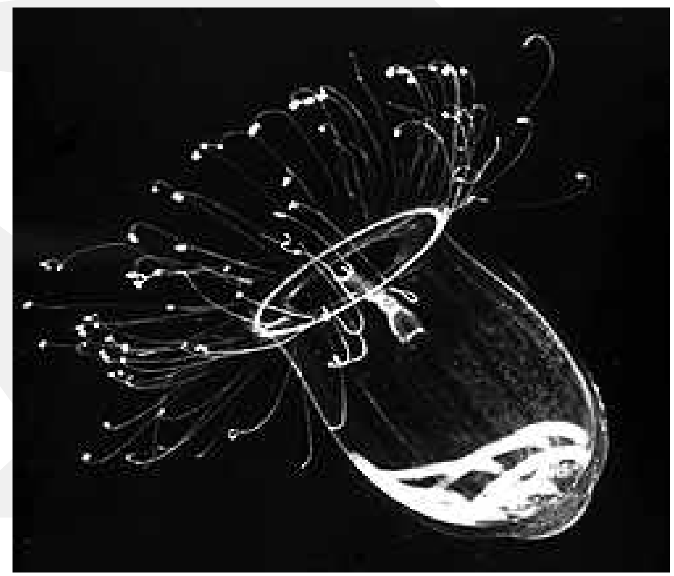
- An *old saying*: “Ontogeny recapitulates phylogeny”
  - Ontogeny: The course or stages of development of an organism.
  - Phylogeny: The history of the evolution of an organism.
  - Example: There is a stage of human development where an embryo has gills.
  - More than a little bit of controversy about this saying…
- Phylogeny reflects application of increasing specialization as a means of harnessing complexity.

Phylogeny and Specialization

- The roots of the family tree:
  - Single cell -> colonies -> multi-cell
  - Increasing cellular specialization as organism scale increases

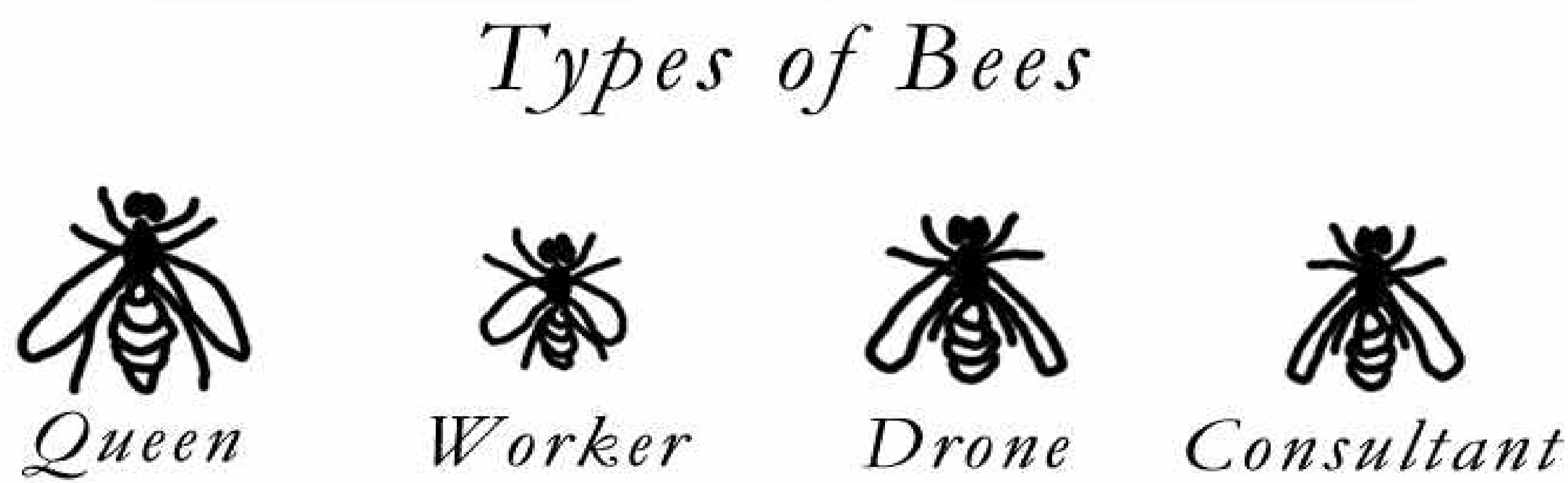






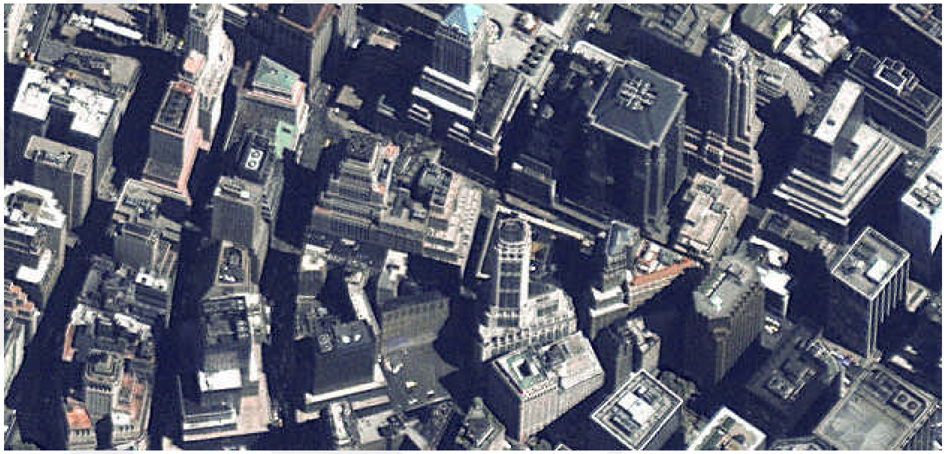
Hierarchical Specialization

- **Insect colony**
  - Specialization of individuals to subdivide work and responsibility
  - Colony acts with higher apparent levels of complexity/capability/intent than individuals.
  - Cost of specialization offset by greater efficiency of the whole.



Societal Specialization

- Specialization enabled human society to rise above agrarian level
  - Government
  - Services
  - Military
  - Business
  - Professional
  - Familial
  - Etc.



- Specialization allows concentration of efforts and resources
- Specialization fosters efficiency

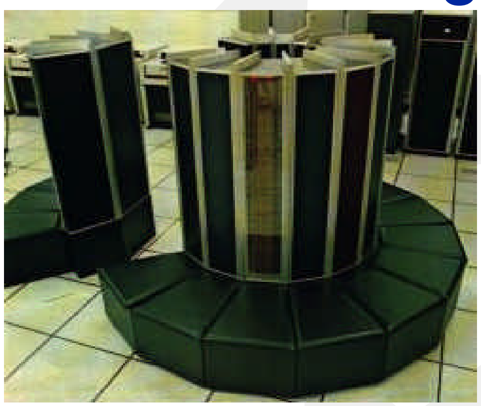
UNLIMITED SCALE, INC.

Specialization Generalizations

- Requires a sufficient plurality of individuals
- Dependencies created between individuals
- Implies prioritization of certain tasks
- Enables coordination of effort to multiply effectiveness
- Assumes execution will benefit from customization of roles

Specialization and Supercomputing

- **Technology specialization diminishing**
- **Is this a problem? Not if your computer room is large enough...**



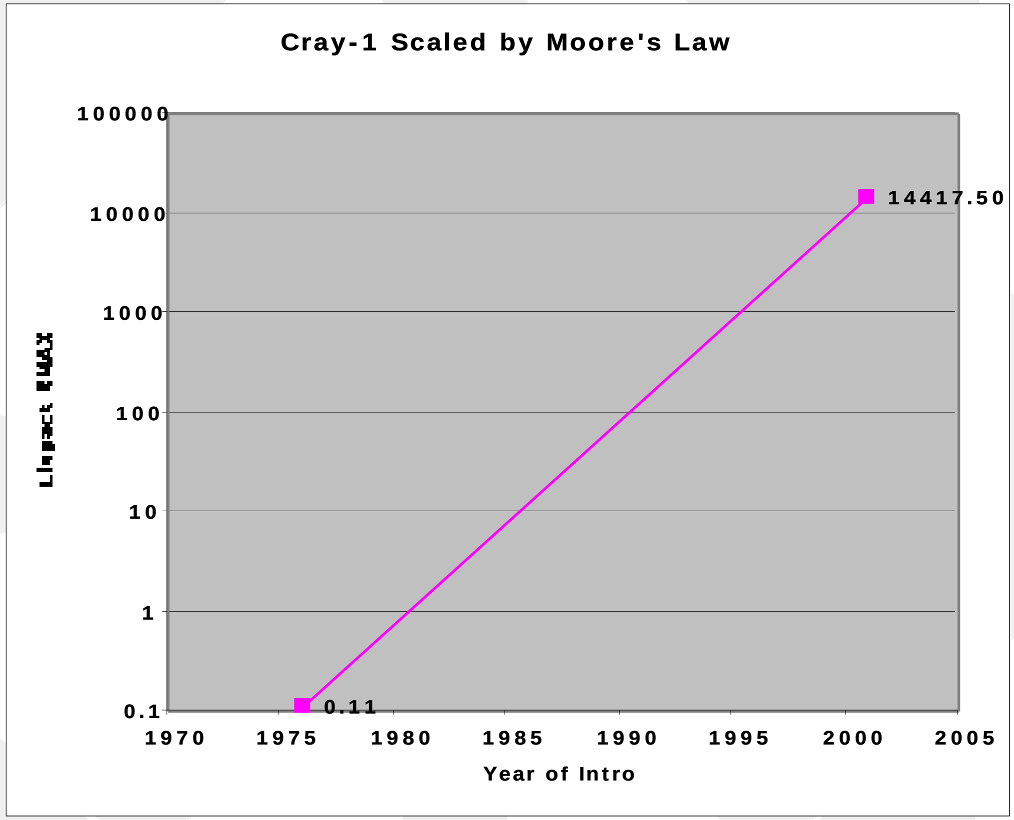
- **Is loss of specialization hurting supercomputing?**

What's a "Supercomputer"?

- **The most powerful computer available. "Super" means *superior, outstanding*.**
- **Example: The Cray-1 in 1976**
  - Specs: 1 CPU, 160 MFLOPS peak
  - LINPACK: ~110 MFLOPS (1K x 1K)
- **Example: ASCI White in 2001**
  - Specs: 8192 CPUs, 12288 GFLOPS peak
  - LINPACK: 7226 GFLOPS (RMAX)

UNLIMITED SCALE, INC.

Moore’s Law and Supercomputing



- Supercomputers today should be demonstrating 14+ TFLOPS RMAX and cost less than $50M (close enough…)

"The TOP500 Supercomputer Sites"

- **June 2001 top dogs:**

1:	IBM	ASCI White, LLNL	7226
2:	IBM	SP Power3, LBLL/NERSC	2526
3:	Intel	ASCI Red, Sandia	2379
4:	IBM	ASCI Blue-Pacific, LLNL	2144
5:	Hitachi	SR8000, U of Tokyo	1709

- **June 2001 bottom of the list:**

496:	HP	Superdome, U of Oslo	68.7
497:	IBM	SP PC604e, BASF	68.5
498:	IBM	SP PC604e, Dregis	68.5
499:	IBM	SP PS2C, CINES	67.8
500:	IBM	SP Power3, Adam Opel AG	67.78

***2 orders-of-magnitude** difference between top and bottom of list*

UNLIMITED SCALE, INC.

Basic Calibration

- "Order-of-magnitude"
  - Speed:
    - Crawling (.6 MPH)
    - Jogging (~6 MPH)
    - Driving (~60 MPH)
    - Flying (~600 MPH)
  - Delay:
    - Come back tomorrow (600 minutes)
    - Go get lunch (60 minutes)
    - Go get coffee (6 minutes)
    - Watch cursor blink (.6 minutes)
- 2 orders of magnitude is a lot.

The Vanishing Supercomputer?

- Over 200 XMPs sold
  - The least powerful was 1/4 of the most powerful.
  - Today's TOP500 #5 is less than 1/4 of #1.
    (Nov. 99: #10)
- Over 200 YMPs sold
  - The least powerful was 1/8 of the most powerful.
  - Today's TOP500 #15 is less than 1/8 of #1.
    (Nov. 99: #32)
- Is there really a problem, or is something else going on?

Possible Explanations

- "Large-scale simulation is less and less useful each year."
  - A "threshold" of capability has been reached.
  - Smaller systems are more than adequate.
- "Parallelism is too hard."
  - "My application only scales to 100 processors."
  - Scientists/engineers are waiting for fully-automatic parallelizing compilers.
- "Supercomputers are too hard to build."
  - "My computer room is too small."
  - On-site integration too painful.
- The "Dark Matter" theory:
  - There are lots of unreported supercomputers, concealed to preserve competitive advantage.

More Likely Explanation

- Supercomputers just don’t work as well as they used to.
  - Computation-to-communication ratio way off, no sparse capability.
    - This may not be insufferable, but will take time.
  - I/O rates haven’t scaled with peak performance and memory sizes.
  - System resiliency and reliability too low to support long runs, demanding user community.
  - Primitive system administration tools make it difficult to integrate into professional production environment.
  - Resulting low overall system efficiency and utilization hurt ROI.
  - Too small fraction of users able can effectively use entire system to justify investment.
    - May as well buy several smaller systems...

UNLIMITED SCALE, INC.

Clusters and Supercomputing

- "We appear to be entering an era of super-computing mono-culture." (Bell/Gray)
- Clusters represent 90% of Top500.
- Even the vector systems (the other 10%) are clustered.
  (I'm going to ignore these.)

What’s Great About Clusters

- Clusters are cheap.
  - Price-performance translates directly to performance in parallel processing world
  - Off-the-shelf components mean faster time-to-market, “fresher” processors.
- Clusters are expandable.
  - Easy to add more nodes, especially if you aren’t trying to scale communications performance.
- Clusters can have fewer single-point-of-failures
  - Assumes you aren’t counting on the whole system to be healthy.
- Clusters can be heterogeneous
  - Seems to be more of a theoretical advantage so far...

UNLIMITED SCALE, INC.

What’s Not So Great (Yet)

- **Clusters don’t really scale very well in a demanding production environment.**
  - Poor computation-to-communication ratio, worse sparse capability.
  - Low I/O performance relative to peak performance and memory sizes.
  - Poor system resiliency and reliability.
  - Primitive system administration tools.
  - Low overall system efficiency and utilization.
- **All these issues related to loss of specialized HW and SW for clustered supercomputing.**
  - Some are difficult to address without affecting price-performance.

UNLIMITED SCALE, INC.

HW Issues

- Communications wish list:
  - Improved fine-grain capability
  - Improved bandwidth
  - Improved latency
  - Integrated synchronization
  - Cheaper
    - "High performance" cluster interconnects cost as much as nodes
- Communications solution:
  - Integrate NIC, router on server nodes
    - Preferably on processor
    - Use put/get model, not I/O DMA model
  - See EV7, only not like EV7.
  - Is Intel listening?
- The other "commodity" HW problems are far less significant.

System SW Issues

Increasing difficulty

- **Integrated system management**
  - Configuration and administration
  - Resource management
  - Accounting and limits
  - Batch job management
  - Scheduling
  - Data center security
- **High-performance I/O**
- **Reliability, resiliency, reconfiguration**
  - Checkpoint/restart
  - Job migration

UNLIMITED SCALE, INC.

Node Specialization

- Specialization needed to increase efficiency, scaling in a cluster

- Some examples:
  - Beowulf -> Scyld
  - Cray T3E
  - ASCI Red and Cplant

Beowulf

- 1994 NASA 16 processor 486 system
  - 1997: Beowulf wins Gordon Bell award for performance/price
  - "Personal supercomputer"
- All nodes peers
  - Though typically a "head" node, I/O node
- Beowulf directions
  - Scyld – master processor and slaves
  - Possible multiple masters in the future

UNLIMITED SCALE, INC.

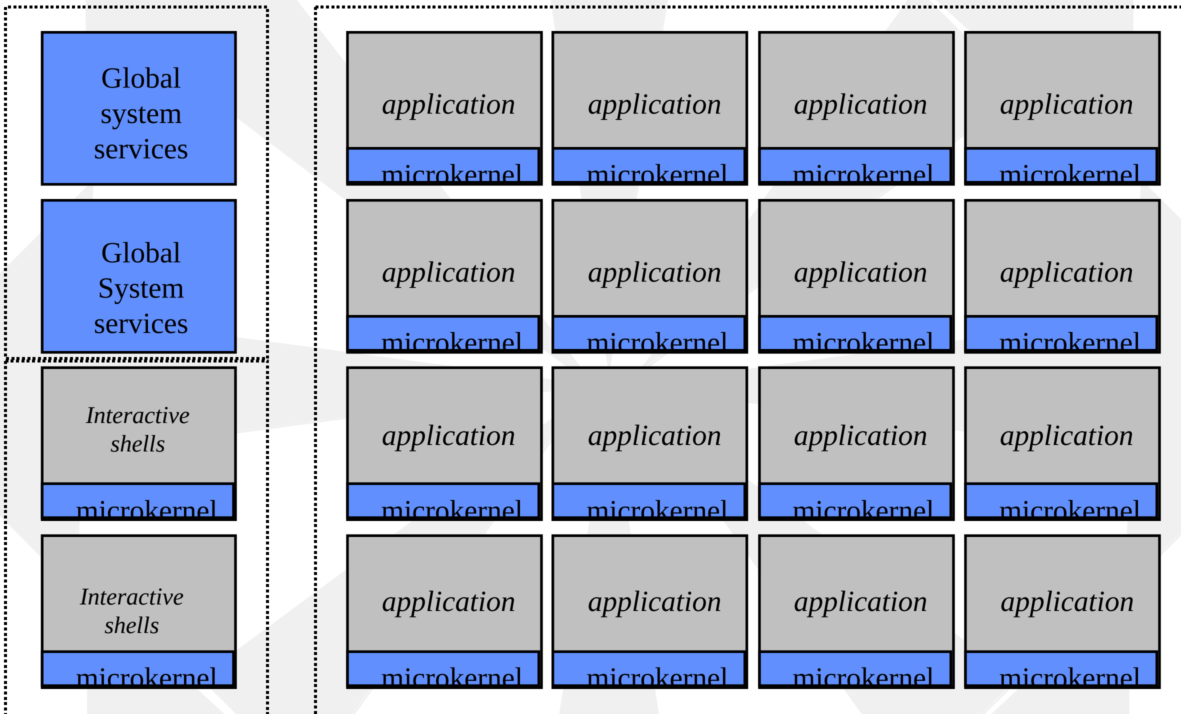
Cray T3E UNICOS/mk

Circa 1996

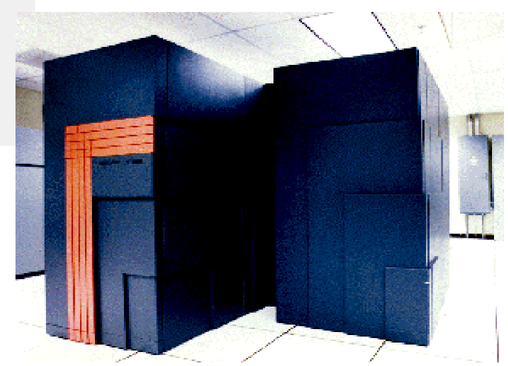
System Nodes

Command Nodes

I/O Nodes (not shown)



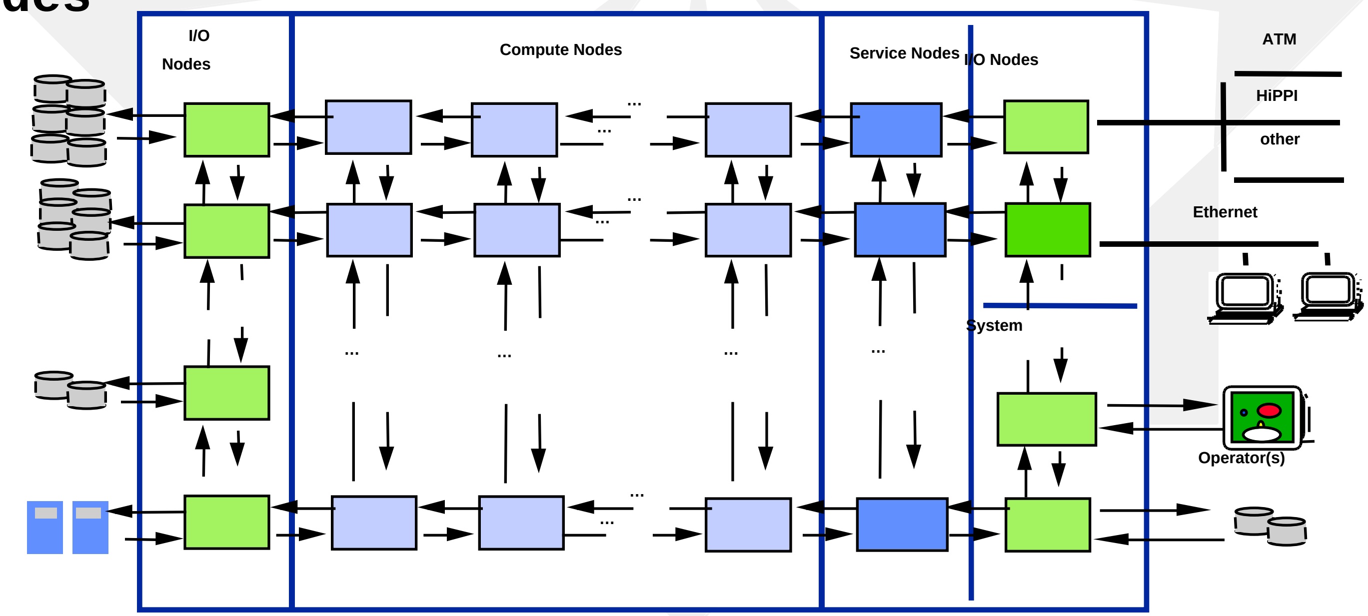
Application Nodes



UNLIMITED SCALE, INC.

Sandia “Microkernel” Work

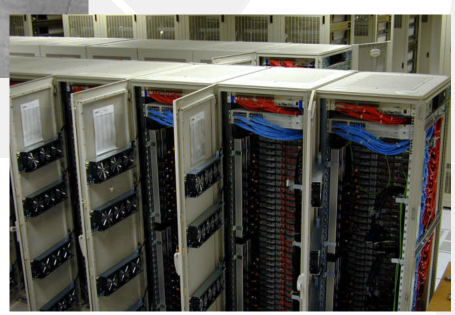
- ~1993 Intel Paragon
  - Originally delivered with OSF1/AD
    - Used most of node memory and was slow
  - Moved to SUNMOS kernel on compute nodes
    - Reorganized system into service, I/O, system, and compute partitions
- ~1996 ASCI Red
  - ~4K nodes and 9K Intel Pentium Pro Processors
  - Puma/Cougar (successor to SUNMOS) kernel on compute nodes

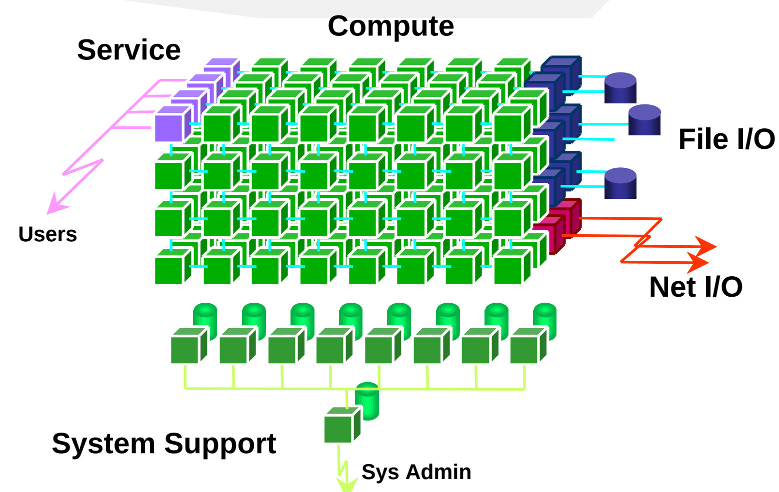


UNLIMITED SCALE, INC.

Sandia and Cplant

- **1997 Cplant**
  - Commodity based
  - "Cluster" using Linux
  - Scales to >1500 processors
  - Brings forward system architecture of specialized nodes from ASCI Red
    - Service, I/O, Compute, and Support





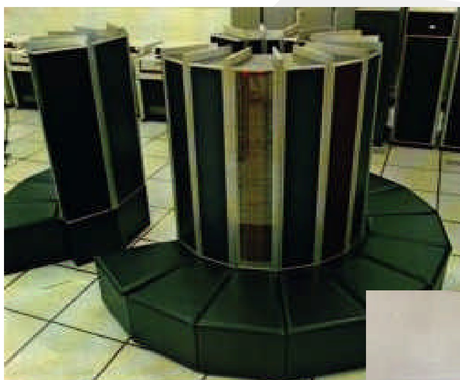
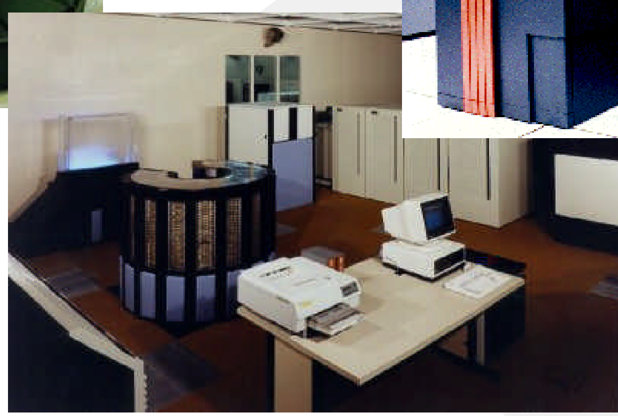
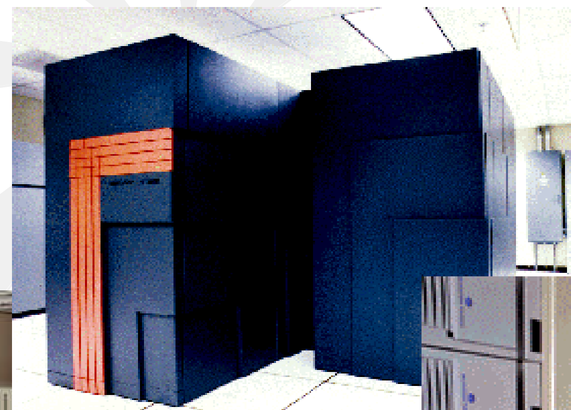
UNLIMITED SCALE, INC.

Unlimited Scale, Inc. Effort

- Create Linux-based scalable system
  - Start with Cplant
  - Drive toward T3E-style SSI
- Lightweight Linux on application nodes
  - Suppress spurious activities -- subject to global control
    - E.g. init, network daemons
- Separate system,I/O node partitions
  - All application processes controlled from system nodes
  - Dedicated I/O nodes, maintenance nodes
- Configuration-driven specialization
  - Nodes “told” at boot whether they are system, applications, I/O or maintenance.

Phylogeny of Supercomputing

- Single processor -> SMP -> cluster
- Custom processor ->
  custom node/standard processor ->
  standard node


UNLIMITED SCALE, INC.

Clusters and Specialization

- *Requires a sufficient plurality of individuals*
  - Nodes available at least possible cost
- *Dependencies created between individuals*
  - Applications nodes, system nodes, I/O nodes, maintenance nodes
- *Implies prioritization of certain tasks*
  - Delivery of cycles to application
- *Enables coordination of effort to multiply effectiveness*
  - System nodes coordinate applications, I/O
- *Assumes execution will benefit from customization of roles*
  - Improved system utilization, higher performance

UNLIMITED SCALE, INC.

Challenges and Opportunities

- **Production-quality supercomputer-class capability must be the goal**
  - “Personal supercomputers” not enough to drive science, engineering, design, discovery.
  - 2 orders of magnitude is nearly a decade.
- **System software is the remaining degree of freedom for architects**
  - Touch the HW, raise the cost.
  - This makes supercomputing harder for everyone.
  - Node specialization works.
- **The real burden is on applications and algorithms.**

Unlimited Scale, Inc.

The Future

- At least another decade of Moore’s Law
- Node prices drop below $1K
- Processor counts rise for all scales
  - Supercomputers: O(100K) processors
  - Personal supercomputers: O(1K) processors
  - Clustered desktops eventually appear
- Intel absorbs networking onto their silicon.
  - Nobody likes it, but it’s too cheap to supplant.
- System software matures.
  - T3E environment equivalence in 5 years.
- Applications lag, but make surprising strides in latency tolerance.
  - Necessity is the mother of invention.

Q&A?