

# Computing Future(s)

*Paul Messina*  
*Director of Science*  
*Argonne Leadership Computing Facility*

---



# Overview

- Evolution of HPC systems – and their building blocks
- Coping with the evolution – new models, algorithms, etc.
- Guiding the evolution – co-design
- Flow down benefits

## Entering a(nother) disruptive phase in Evolution (Punctuated Equilibrium \*)

- Hardware technologies are reaching limits that will result in major changes in system architectures and therefore in applications and software; these issues will impact computers of all sizes
  - Although Moore's Law\*\* will continue to hold for at least another decade, it will no longer yield faster clock speeds (assuming CMOS technology)
    - that will force major changes in HPC architectures
  - Most of the speed increase will come from increased parallelism, much of it within a node (think of a node as a chip or a collection of chips on a board, connected to other nodes by a network for data transfers)
    - that will pose programming challenges (applications, tools, system software) and increase in energy use for large systems

*\*punctuated equilibrium/evolution was postulated by S.J. Gould for evolution of life on Earth*

*\*\* Moore's Law: The number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years. The increase in speed was a by-product of this trend.*



# Previous transitions in computing

## ■ Sequential to HPC sequential

- Memory access was an issue (e.g., caches, SCM and LCM)

From Wikipedia, re the CDC 6600, designed by Seymour Cray:

- Unlike most high-end projects, Cray realized that there was considerably more to performance than simple processor speed, that I/O bandwidth had to be maximized as well in order to avoid "starving" the processor of data to crunch. As he later noted,
- *Anyone can build a fast CPU. The trick is to build a fast system.*

## ■ Sequential to vector

- Vectorizing algorithms were needed to achieve high performance, compilers helped

## ■ Vector to parallel, later highly parallel

- Also array processors (precursors to GPUs)
- Many issues in programming, I/O, algorithms, macro architecture



# Technology trends

- The biggest systems in 2020 will have O(100M) cores, O(1B) threads
- The biggest increase in parallelism will be within a node
  - Nodes with many cores/processors, say 1,000
  - Chips with 100s of cores
- Flops to memory size ratios will increase due to power and parts costs
- Memory speed will not match processor speeds
  - Most memory will be remote, even within a node, and distributed
    - It is expensive for memory to be shared among many processors
  - This is a continuation of a 20+-year trend but ratio is getting really unbalanced.
- Commodity architectures and components will be the building blocks for the foreseeable future
  - With some tweaks for us (scientific computing)



# Applications are evolving too

- Multi-physics: integrated codes that combine multiple models to simulate a phenomenon (e.g., climate, combustion)
  - Physics, chemistry, engineering
- Uncertainty quantification
- Data-intensive
- All of the above
  
- *That evolution adds complexity even if computer architectures did not change*

# Consequences

- Have to face parallelism and concurrency in most (all?) systems, regardless of size
  - Need to find ways to exploit all that parallelism
- The number of components in the biggest systems will result in more frequent faults
  - Applications may need to become more resilient
- Flops are “free,” memory is expensive
  - All those flops present an opportunity
- Integrated codes lead to additional load balancing challenges
  - Another task for application developers

## Key Issues: Accessing, moving, and storing data

- Memory size and access time have long been recognized as a key aspect of the configuration/architecture
- Need to reduce data movement; it is expensive in time and power consumption
  - Latency to access remote memory will be  $O(1,000)$  vs. on-chip
    - And most memory will be remote
  - Power consumption for on-node memory access will be  $\sim 20x$  that for on-chip, off-node  $\sim 50x$
  - Communication-avoiding algorithms will help
  - More integrated memory and functional units within chip will help
- I/O has been a bottleneck “forever”
  - With highly parallel machines it is worse
- Similar concerns about storage and data analysis



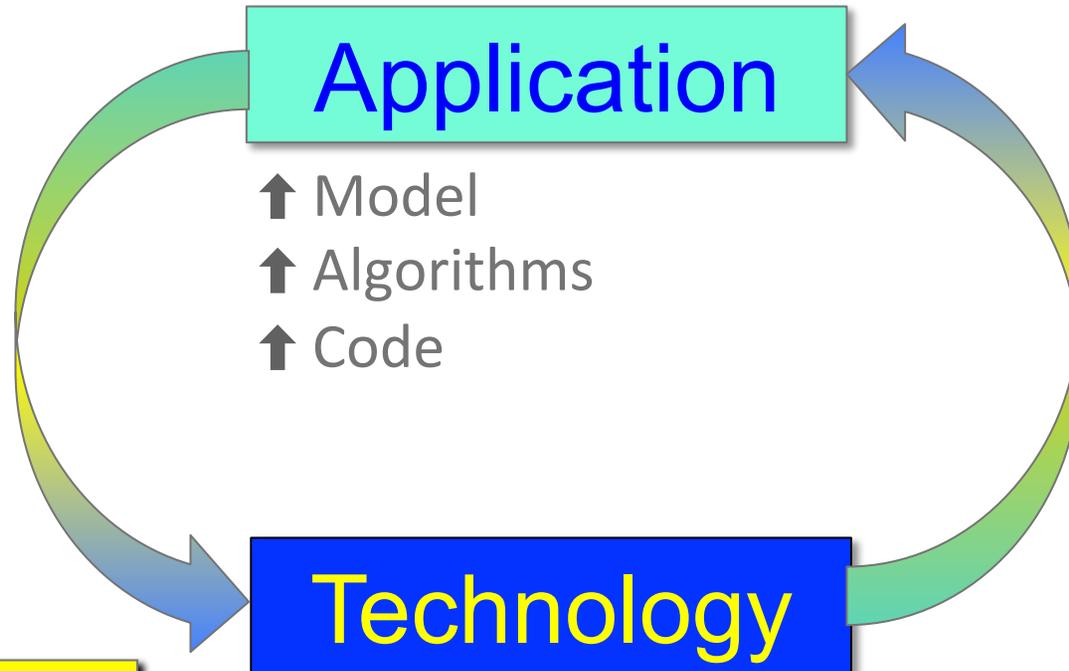
# We can guide the evolution - we have throughout HPC history

- Through **co-design** we can influence the evolution
  - up to a point
  - the technologies are more complex
  - the applications are much more complex
- Mathematical models, programming models, algorithms, etc. are an integral part of the co-design effort
- The existence of issues that arise in applications from many domains – **cross-cutting issues** – means that by working together and using a co-design and co-development approach, the resulting systems (hardware and software) will be better matched to our applications than would otherwise be the case



# Co-design leads to systems that are better suited for science

Application driven:  
Find the best  
technology to run  
this code.  
*Sub-optimal*



*Now, we must expand the co-design space to find better solutions:*

- *new applications & algorithms,*
- *better technology and performance.*

Messina: Computing Futures

Technology driven:  
Fit your application to  
this technology.  
*Sub-optimal.*



## Selected observations gleaned from discussions with ~ 40 computational science teams

- All Code teams:
  - Portability is critical
  - Large, established community and code base
  - Are exploring how to express more parallelism
- Many codes:
  - Hide parallelism in framework
  - Counting on dynamic execution environment for balancing load (tasks, data redistribution, threads)
- Shared worries:
  - How to manage faults
  - Data sizes and management
  - Debugging
  - How to express more parallelism



# High-end systems will be compatible with the rest of the scientific computing ecosystem

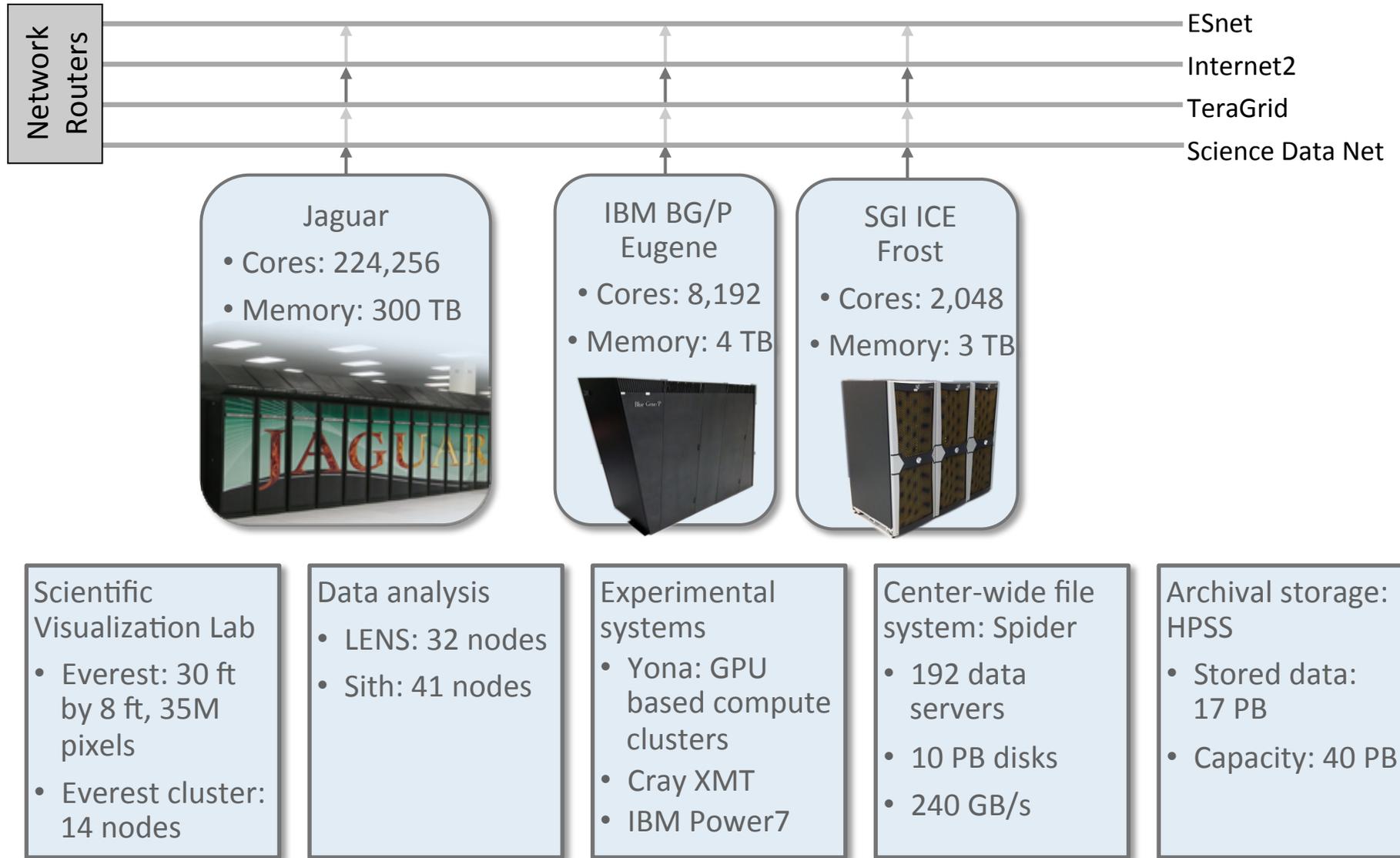
- Cost and sustainability matter
  - Computer generations are short compared to telescopes and particle accelerators
  - **AND CODES**
- To highest-end systems will be similar enough to mid-range and small systems to make portability realistic, up and down the scale (unlike sequential vs. vector, MP/clusters vs. vector)
  - within the architectures that are implemented
  - two likely variants are
    - systems whose nodes use homogenous cores and
    - systems with heterogeneous cores that include specialized processors (e.g., vector/array processors)

# Leadership Computing Facilities

## - testbeds for tomorrow's systems

- LCFs were established at Oak Ridge and Argonne to provide the computational science community with leading-edge computing capabilities dedicated to breakthrough science and engineering
- Support the primary mission of DOE's Office of Science Advanced Scientific Computing Research (ASCR) program to discover, develop, and deploy the computational and networking tools that enable researchers in scientific and engineering disciplines to analyze, model, simulate, and predict complex phenomena that require the power of the LCF resources.

# ORNL resources

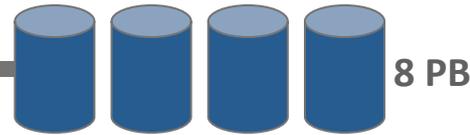


# ALCF resources

## Production

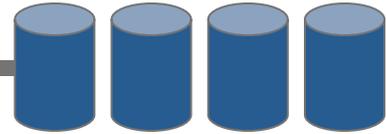
### Intrepid

- 40k nodes
- 160k cores
- 556 TF



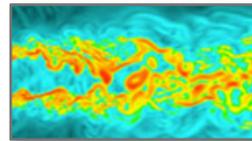
DDN 9900s

8 PB



640 TB

DDN 9550s



### Eureka (Viz)

- 800 cores
- 50 NVIDIA S4 GPUs
- 100 TF



8 PB

Spectra Logic T950

### User teams

- Esnet
- UltraScienceNet
- Internet2

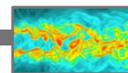
## Test and development

### Surveyor

- 1 rack
- 4k cores
- 13.9 TF



128 TB  
DDN 9550



### Gadzooks (Viz)

- 32 cores, 2 NVIDIA S4 GPUs



# New ALCF Resources Coming Soon

- *Mira* - Blue Gene/Q System
  - 49,152 nodes / 786,432 cores
  - 786 TB of memory
  - Peak flops rate: 10 PF
- New Visualization System
  - State-of-the-art server cluster with latest GPU accelerators
  - Provisioned with the best available parallel analysis and visualization software
- Storage
  - ~30 PB capability, 240GB/s bandwidth (GPFS)
  - Storage upgrade planned in 2015
    - Double storage capacity and bandwidth



# Upgrades of OLCF are coming even sooner

- Rolling upgrade of nodes in a few months
  - 12 => 16 cores
  - New network
- Addition of GPUs a bit later
- Peak ~ 20 petaflops

# Allocations @ LCF

	INCITE <span>60%</span>		ALCC <span>30%</span>		ALCF Discretionary <span>10%</span>	
<b>Mission</b>	High-risk, high-payoff science that requires LCF-scale resources*		High-risk, high-payoff science aligned with DOE mission		Strategic ANL and ASCR use	
<b>Call</b>	1x/year – (Closes June)		1x/year – (Closes February)		Rolling	
<b>Duration</b>	1-3 years, yearly renewal		1 year		3m,6m,1 year	
<b>Typical Size</b>	<b>30 - 40 projects</b>	<b>10M - 100M core-hours/yr.</b>	<b>5 - 10 projects</b>	<b>1M – 75M core-hours/yr.</b>	<b>100s of projects</b>	<b>10K – 1M core-hours</b>
<b>Review Process</b>	<b>Scientific Peer-Review</b>	<b>Computational Readiness</b>	Scientific Peer-Review	Computational Readiness	Strategic impact and feasibility	
<b>Managed By</b>	INCITE management committee (ALCF & OLCF)		DOE Office of Science		LCF management	
<b>Availability</b>	Open to all scientific researchers and organizations * <b>Capability &gt;20% of cores</b>					



# Questions?