Sandia National Laboratories

SPECTRA

Turning the Titanic with a Leaf-blower:

How the DOE, Influences, Leverages and Adapts to an Evolving HPC Industry

ADVANCED ARCHITECTURE TESTBED PROGRAM

VANGUARD

James H. Laros III

Distinguished Member of Technical Staff

Dept. 01422, Scalable Computer Architectures

Trilinos User Group (TUG) 2023

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Request ID 1726093

Abstract

In this presentation, I will give a *hopefully* informal **interactive** talk attempting to touch on various topics related to the history of High-Performance Computing (HPC) in Sandia's Center for Computing Research (CCR) and what the future might hold. As we gaze into the crystal ball together, we will look at the motivations and possibilities for future technologies and speculate on what we may see in the coming 5-10 years, while avoiding Non-Disclosure Agreement (NDA) information. I'll touch on the NNSA platform strategy and present some history of the path from prototypes to large platform installations using examples from first our Advanced Architecture Test Bed program and more recently our Vanguard program and the impact we hope to have on future NNSA architectures.

Note: An additional presentations is included with this deck, previously presented history of HPC in center 1400 for wide audience and specifically for Postdoc and Early Career Seminar Series for your reference.

Overview

Brief History of HPC in center 1400

Impactful Sandia programs

- Advanced Architecture Testbed Program
- Vanguard Program
- DOE *Forward like Programs

Non-NDA Crystal ball ☺

History of HPC in center 1400

- The DOE Labs role in HPC cannot be overstated
- Center 1400* since its inception has had a huge impact on HPC in many areas
  - Hardware
  - Software
  - Programming Models

Paragon
1993

ASCI Red
1996

Cplant
1998

Red Storm
2005

ADVANCED ARCHITECTURE
TESTBED PROGRAM

Advanced
Architecture
Testbed Program
2010 - present

VANGUARD

Astra
2018

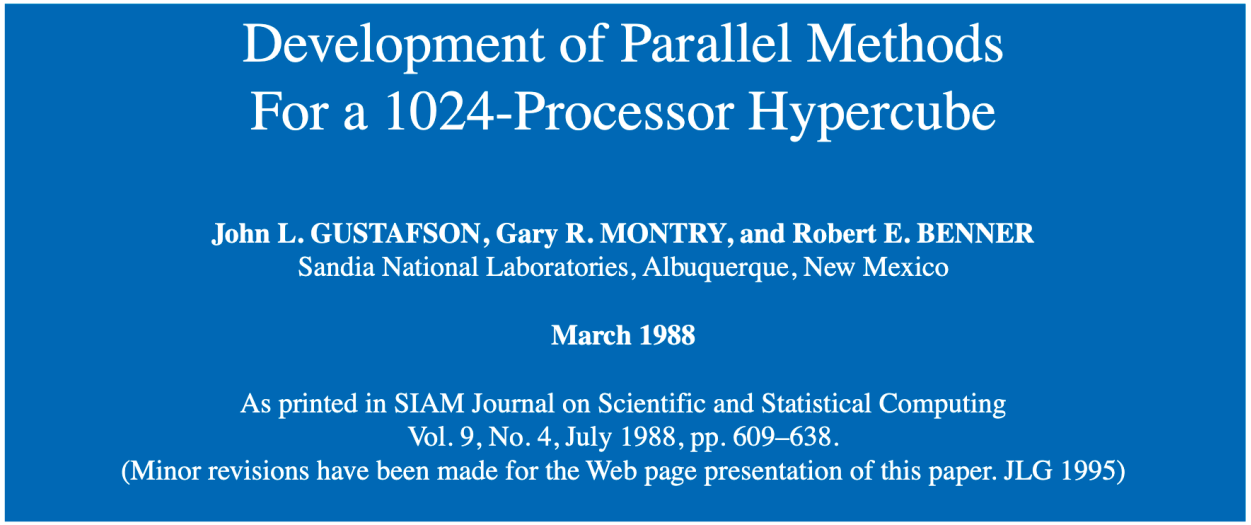
VANGUARD

SPECTRA

2025

Note, large number of background slides provided for your reference

Preamble



- 1988
  - Karp Award for demonstrating unprecedented speedups using processors working together compared to processors running separately
  - 1st Gordon Bell Prize for achieving a thousandfold speedup on three engineering problems analyzed with 1,024 processors working in parallel.
  - **Inspired a sea change in how Sandia and the world approached scientific computing**

History of HPC in center 1400 (cont.)

Paragon
1993

ASCI Red
1996

Cplant
1998

- Intel Paragon
  - Sandia's first #1 on Top500 in 1994
  - 2nd Gordon-Bell Prize - *Applications of Boundary Element Methods on the Intel Paragon*
  - Portals version 0 light-weight Operating system

Cougar

portals

- ASCI Red
  - #1 on Top500 from June 1997 to June 2000
  - First TeraFLOPs computer, first production MPP
  - Sandia Architecture, Cougar, Portals version 2, significant contributions to MPI and scalability
- Cplant - Sandia kicked off the island ☹
  - Adapted – Emulated ASCI Red software environment
  - Largest clusters by far at this point in time
  - Reached #30 on Top500 – Self Made Supercomputer
  - Linux makes its debut
  - Possibly most importantly, prepared 1400 for Red Storm

CPLANT
Computational Plant

History of HPC in center 1400 (cont.)

Red Storm
2005

- Red Storm – We'll make our own island ☺
  - Sandia designed Architecture
  - Marriage of commodity and proprietary
    - AMD Opteron first 64 bit x86 processor
    - Seastar high-speed interconnect
  - Portals version 3
  - Cougar is now Catamount
  - Reached #2 on Top500
  - This effort saved Cray from certain doom
    - **"Without Red Storm I wouldn't be here in front of you today. Virtually everything we do at Cray — each of our three business units — comes from Red Storm. It spawned a company around it, a historic company struggling as to where we would go next. Literally, this program saved Cray." – Peter Ungaro Cray CEO**
    - Worked out pretty well for Sandia also
  - Most successful SuperComputer line EVER
  - Bill Camp - Recipient of 2016 IEEE Computer Society Seymour Cray Computer Engineering Award "for visionary leadership of the Red Storm project, and for decades of leadership of the HPC community."
  - Frontier and El Capitan direct decedents of Red Storm

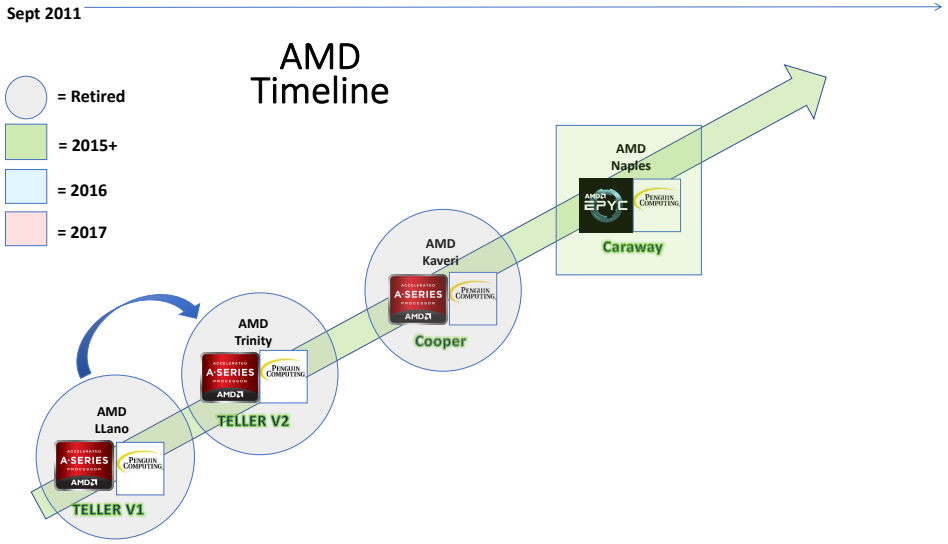
**Note the bytes/FLOP ratio in the table we will get back to this later.*

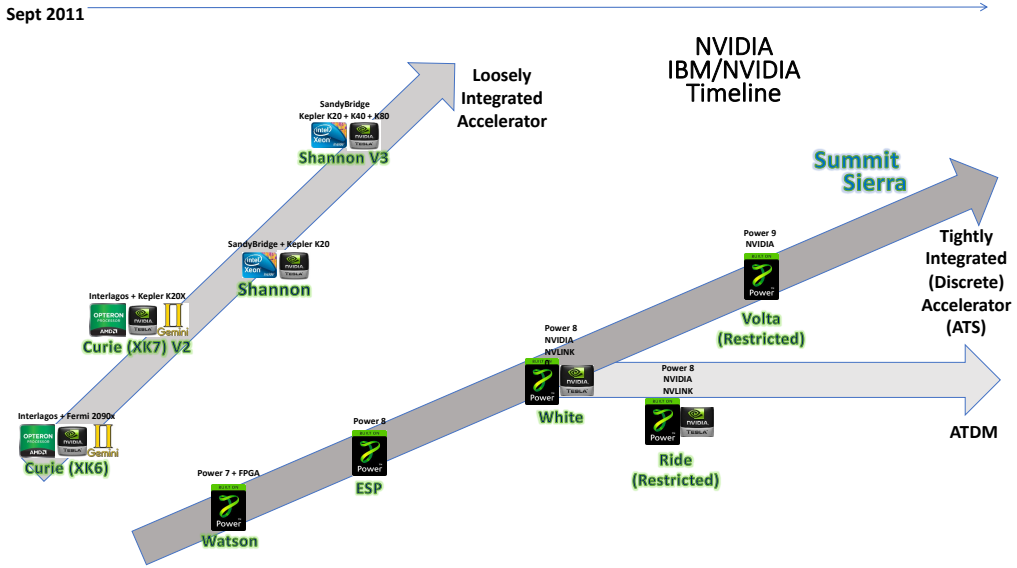
SNL Red Storm / Cray XT3
2.0 GHz AMD Opteron Single-Core ~2004
57.99 TB/sec BW / 43.52 TF Peak = **1.33 B/F**
Network B/F approximately 0.6
2.4 GHz AMD Opteron Dual-Core
78.14 TB/sec BW / 130.56 TF Peak = **0.6 B/F**
Network B/F approximately 0.5 B/F

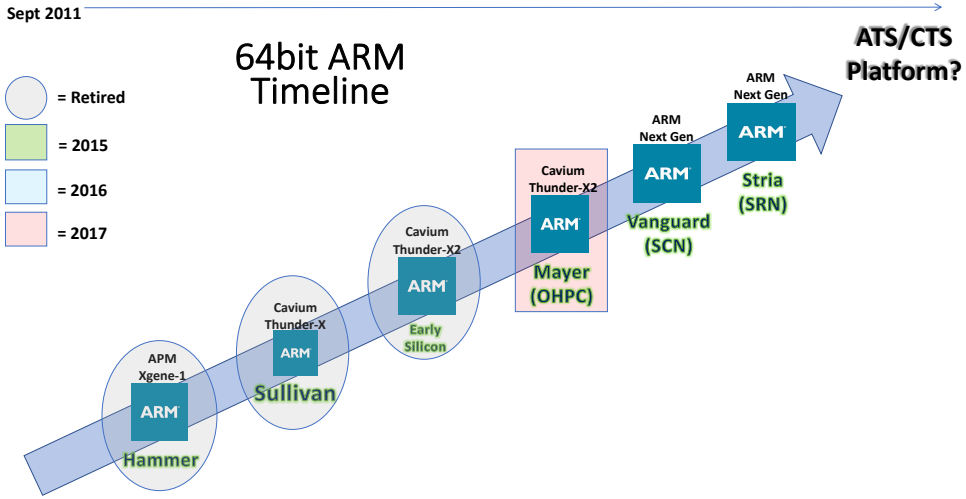
History of HPC in center 1400 (cont.)

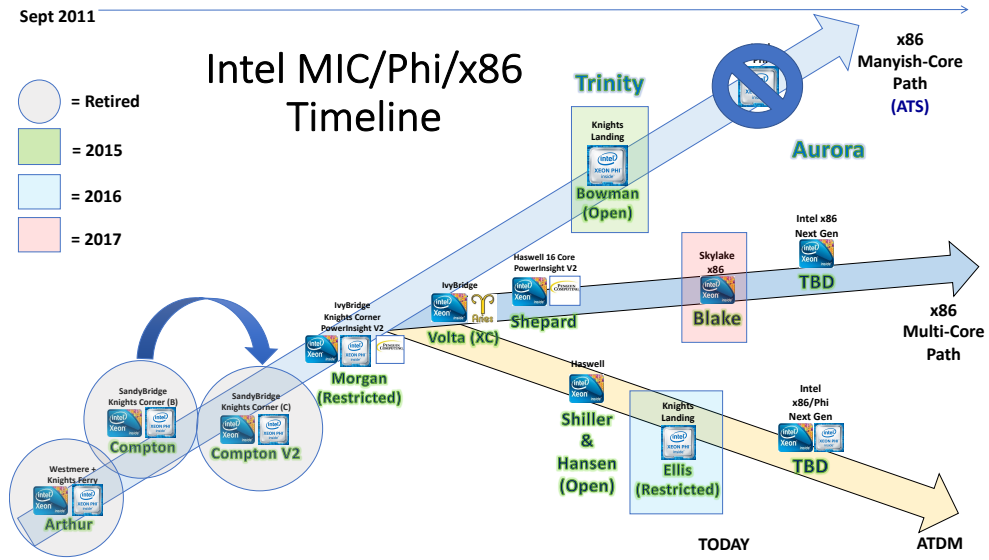
ADVANCED ARCHITECTURE TESTBED PROGRAM

- Explosion of new node architectures emerging
- In the absence of large-scale deployments how can we have impact?
- Broadly focused, node to rack scale investigations
- Developed lasting partnerships with a wide variety of technology providers and integrators
- Leveraged Mini-apps to provide feedback to partners regarding Tri-lab application requirements
- Motto – to be a scout for future computer architectures


Sept 2011
AMD Timeline
= Retired
= 2015+
= 2016
= 2017
AMD LLano
TELLER V1
AMD Trinity
TELLER V2
AMD Kaveri
Cooper
AMD Naples
Caraway

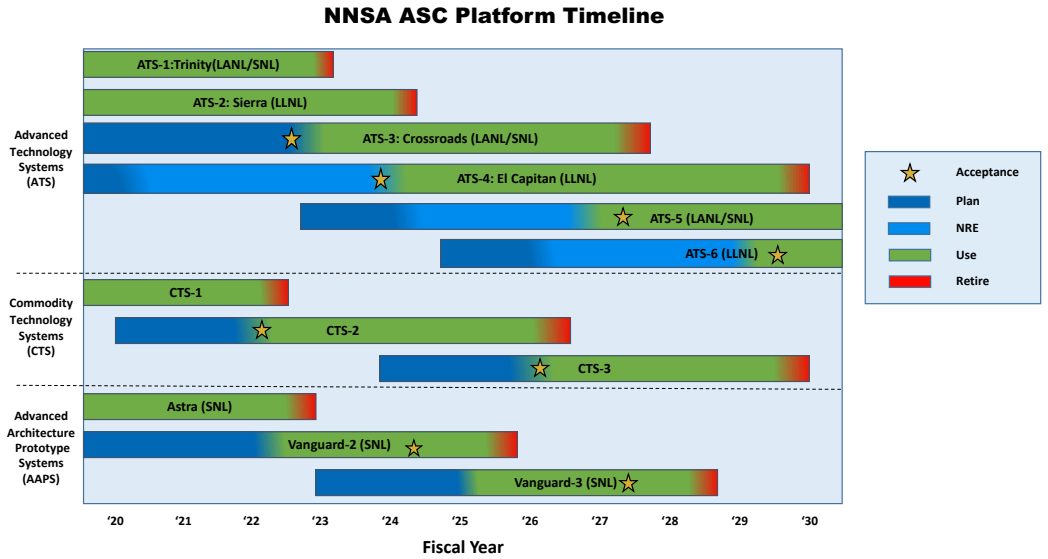

Sept 2011
NVIDIA IBM/NVIDIA Timeline
Loosely Integrated Accelerator
Tightly Integrated (Discrete) Accelerator (ATS)
SandyBridge Kepler K20 + K40 + K80
Shannon V3
SandyBridge + Kepler K20
Shannon
Interlagos + Kepler K20X
Curie (XK7) V2
Interlagos + Fermi 2090X
Curie (XK6)
Power 7 + FPGA
Watson
Power 8
ESP
Power 8 NVIDIA NVLINK
White
Power 9 NVIDIA
Volta (Restricted)
Summit Sierra
Power 8 NVIDIA NVLINK
Ride (Restricted)
ATDM


Sept 2011
64bit ARM Timeline
ATS/CTS Platform?
= Retired
= 2015
= 2016
= 2017
APM Xgene-1
Hammer
Cavium Thunder-X
Sullivan
Cavium Thunder-X2
Early Silicon
Cavium Thunder-X2
Mayer (OHPC)
ARM Next Gen
Vanguard (SCN)
ARM Next Gen
Stria (SRN)

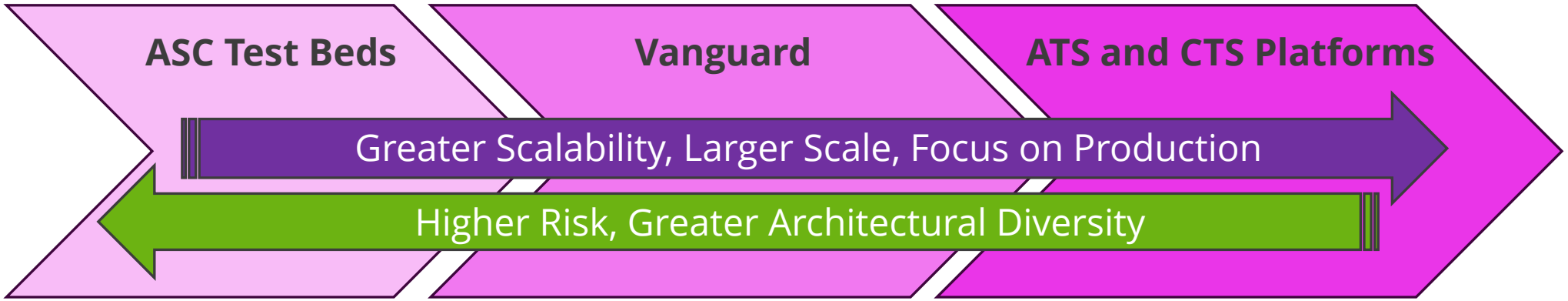

Sept 2011
Intel MIC/Phi/x86 Timeline
= Retired
= 2015
= 2016
= 2017
x86 Manyish-Core Path (ATS)
Trinity
Knights Landing
Bowman (Open)
Aurora
IvyBridge Knights Corner PowerInsight V2
Morgan (Restricted)
IvyBridge
Volta (XC)
Haswell 16 Core PowerInsight V2
Shepard
Skylake x86
Blake
Intel x86 Next Gen
TBD
x86 Multi-Core Path
SandyBridge Knights Corner (B)
Compton
SandyBridge Knights Corner (C)
Compton V2
Westmere + Knights Ferry
Arthur
Haswell
Shiller & Hansen (Open)
Knights Landing
Ellis (Restricted)
Intel x86/Phi Next Gen
TBD
TODAY
ATDM

History of HPC in center 1400 (cont.)

- Sandia chosen to host the first Arm-based supercomputer
  - The return of large-scale platforms at Sandia
- Huge amount of pressure and a compressed timeline
  - Feels like Red Storm all over again
- Advanced Architecture Prototype Program (Vanguard) is born as a result of Sandia's success
  - **Goal: prove the viability of emerging advanced architecture technologies for the NNSA mission**
- 725 West hosted first x86_64 based Supercomputer – Red Storm
- 725 East will host first Arm64 based Supercomputer – Astra
- Astra did lead to a selection of an Arm-based processor for



February 2021



~~History~~ Current status of HPC in center 1400 (cont.)

- Now we run into NDA issues
- Advanced Architecture Testbed program is alive and well
  - Intentionally blurring the line between the testbed program and Vanguard when it makes sense
- Sandia is partnering with a start-up to deploy a novel accelerator
  - More information announced at SC23
- Goal remains the same, prove the viability of emerging advanced architecture technologies for the NNSA mission
  - This one is definitely novel
- We are accepting more risk in hopes of greater impact to the program
- Given Sandia's history we feel this is our role
- We have been pursuing heterogeneous platform topics and tightly integrated architectures for many years now

VANGUARD

SPECTRA
2025

ADVANCED ARCHITECTURE
TESTBED PROGRAM

Crystal Ball

- Any look into a crystal ball would have to include the topics of memory bandwidth and memory latency
- Historically, our applications have been memory bound (bandwidth and latency).
  - Note improving latency is much harder, pesky speed of electrons considerations
  - Can be somewhat mitigated by employing parallelism as can bandwidth
  - Only gets us so far
- **Recall the bytes/FLOP ratio mentioned for Red Storm (B/F 0.66)**
- **Frontier has a bytes/FLOP ratio of 0.03**
  - **22x worse than Red Storm**
- The memory bottleneck that has grown contributes to a huge lack of efficiency for our applications on current architectures
  - As low as single digit percentages relative to theoretical peak
- The introduction of HBM has given us an improvement in memory bandwidth that we have not seen since the memory controller was integrated into the ASIC almost 20 years ago
  - Still not enough as indicated by the B/F ratio of Frontier
  - So what are we doing about this?

Crystal Ball

- Advanced Memory Technology Program
  - Started in 2021
  - Two rounds of funding executed thus far, next slide
- Vanguard III in the early days of planning but looks very promising in a number of dimensions, especially memory bandwidth and latency
  - Progress with B/F ratio?
- Industry wise, advanced packaging techniques are entering into production chips
  - 2.5D and 3D packaging
  - Chiplets
- Additional advanced packaging techniques that will enable tighter integration and increased processor to memory and processor to processor
  - Chiplet, 2.5D or 3D stacked technologies here and on the horizon (details largely are NDA)
- DOE *Forward like investments in storage, networking and other topics will help us effect future architectures
  - Some announcements coming at SC23

Advanced Memory Technology Program (Tri-lab, Sandia led)

Advanced Memory Technologies

- New materials for memory
- Novel foundational DRAM designs
- Compute/Memory Architecture

Goal to achieve an NNSA application performance improvement of 40X

Three awards started in 2022/23:

1. Cerebras - Demonstrate potential NNSA application performance gains by mapping, implementing and optimizing **NNSA motifs on current and future Cerebras wafer-scale HW.**
2. AMD - **Investigating advanced 3D stacked memory** technology approaches to significantly increase memory bandwidth, optimize and enhance near-memory data manipulation and marshaling for a range of processor types.
3. Intel – **Investigating low power, high density, high bandwidth memory systems** that can deliver sustained efficiency for sparse and dense workloads.

Co-design activities with vendors are critical for ensuring future needs are met (*Forward previously mentioned for example)

Discussion

Sandia National Laboratories

Exceptional service in the national interest

Those who fail to learn from history are condemned to NOT repeat it

PRESENTED BY

JAMES H. LAROS III

DISTINGUISHED MEMBER OF TECHNICAL STAFF, DEPARTMENT 01422,
SCALABLE COMPUTING ARCHITECTURES

SAND2021-7102PE

Before we get started

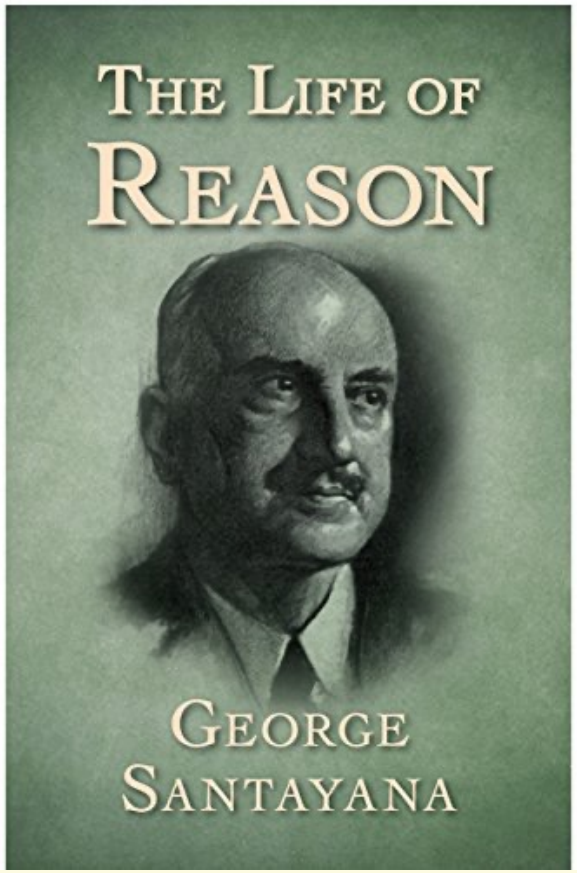
Contributions to the material used in this presentation was provided by

- Bill Camp
- Sue Kelly
- Jim Tomkins
- Rolf Riesen
- John Noe
- Kevin Pedretti
- Steve Monk
- Ron Brightwell
- Google

All played significant roles in the projects presented along with many more not listed

including Google for mining information ☺

Note: this is a lot of ground to cover, each effort would take hours to cover individually. Excuse me in advance for leaving out significant details and events.



National Churchill Museum - https://www.nationalchurchillmuseum.org/blog.html

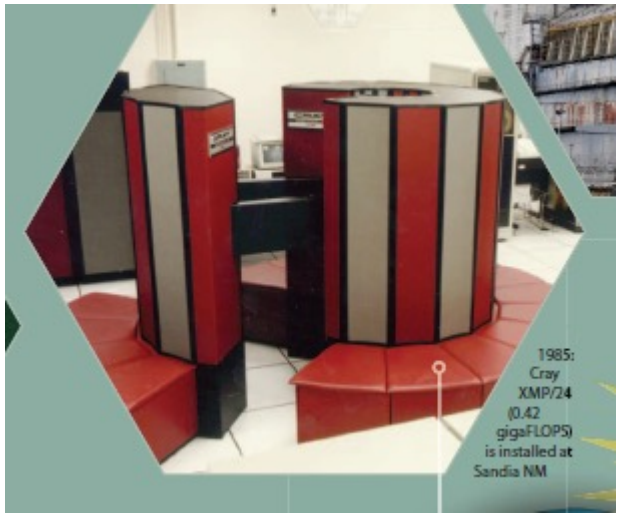
Why do we care about history

- Lots of answers to this one
- I will claim, we want to understand what worked well in the past and replicate it in context with the current environment
  - What didn't work well is also an important lesson
- I'll try and elaborate by reviewing some high points in, mostly, our centers, HPC history
  - I tend to speak of these efforts in a positive way but I think accurately
  - Important things happen outside our center sometimes ☺
- I'll cover hardware and software, in my opinion both are essential parts of the equation
- I'll try to avoid stories about walking miles to school, uphill in both directions in the snow
  - Although I did this but not uphill both ways ☺
- I'll also try to make this somewhat entertaining
  - Warning, my sense of humor can sometimes require explanation
- Anyone that is present for this and participated is welcome to add dimension to what I'll say

Computing at Sandia - Early days

- Sandia invested in a variety of what are now called commercial off the shelf (COTS) computers starting in the 60's
- There were a wide variety of computer companies through the 1980's with competing offerings
- These machines were sometimes used for a single purpose or a single early code
- Note the dominance of Cray Vector architectures in the 80's

60's	70's	80's
IBM 704 CDC 1604 CDC 3600 CDC 6600 Univac 1108 IBM 7090	CDC 6600 CDC 7600 Cyber 76 MultiMainframe ECS	CDC 6600 Cyber 76 MultiMainframe ECS Cray 1S 500 Cray 1S/1000 Cray 1 Cray X-MP2 CDC 855 NOS Cray X-MP4 CDC 855 NOS/VE Cray Y-MP2 Cray Y-MP8

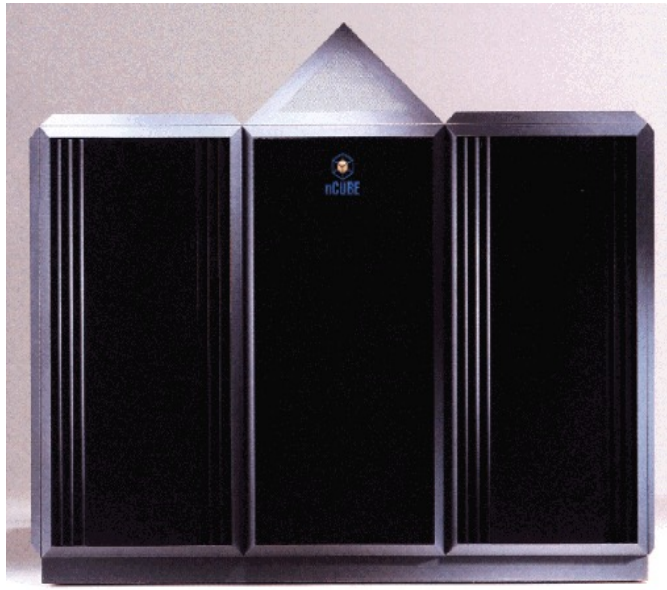


Vector Architectures: Past, Present and Future, Roger Espasa, Mateo Valero, James E Smith 1988
http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15740-f03/public/doc/discussions/uniprocessors/vector/vector-past-present-future-supercomputing98.pdf

Then something happened

- Previously, Sandia's focus was NOT on leading/bleeding edge technology
- Ed Barsis and soon after Bill Camp led the way to exploring what at the time was a pretty drastic alternative, massively parallel processing (MPP) systems
- Sandia fielded exploratory systems from Alliant, Thinking Machines and nCube
- Significant research and development was required to understand how to exploit extreme levels of parallelism and distributed memory
- New parallel programming models were being developed PVM and MPI
  - MPI became the de facto standard of course
- Light-weight operating systems were developed that used only a small amount of scarce node memory resources
  - Leave the memory to the application
- Using these new architectures required a completely new approach
  - Often requiring algorithmic changes in addition to programming model changes

Thinking Machines



nCUBE

PVM - https://www.csm.ornl.gov/pvm/

Looks like we might have something here

Development of Parallel Methods For a 1024-Processor Hypercube

John L. GUSTAFSON, Gary R. MONTRY, and Robert E. BENNER
Sandia National Laboratories, Albuquerque, New Mexico

March 1988

As printed in SIAM Journal on Scientific and Statistical Computing
Vol. 9, No. 4, July 1988, pp. 609–638.
(Minor revisions have been made for the Web page presentation of this paper. JLG 1995)

- Recognized by the Gordon Bell Award and the Karp Prize, at IEEE's COMPCON 1988 meeting in San Francisco on March 2
- Work was accomplished on the largest nCUBE 2 system installed, a 1,024 CPU (1.91 gigaFLOPS) system at Sandia Laboratories
- Ran the nCX microkernel but also SUNMOS (Sandia/UNM Operating System)
  - Beginning of a long line of Light-weight Kernel development at Sandia, more later
- Demonstrated efficient parallel solutions for three full-scale scientific problems
  - Wave mechanics
  - Fluid dynamics
  - Structural analysis
- This is one seminal event that inspired a sea change in how Sandia approached computing
  - I would argue in turn influencing the wider community
- Center 1400 leading the way to adopting large-scale distributed memory MPP architectures

Before we proceed

- So why are we doing this?
- In the early 90's some important activities at the National level
- In 1992, the US conducts its last Nuclear Weapons test
  - It was thought that testing would continue until 1996 but additional testing is never approved
- Work is in progress to establish the DOE stockpile stewardship program
  - 1994 National Defense Authorization Act established the Stockpile Stewardship Program
  - In the absence of testing - we needed a new way of assessing the performance of nuclear weapon systems, predict their safety and reliability and certify their functionality
- The Accelerated Strategic Computing Initiative (ASCI) Program is established
  - Modeling and Simulation will replace testing
- 50 years of nuclear testing will be combined with a new simulation capability to provide high-confidence reliable predictive capability
- Small problem, commercial computing offerings if left on their current trajectory will not meet the projected requirements for modeling and simulation to replace testing
  - Target 2004 – 2010 to have usable working ASCI computer systems and application simulation capability
    - Huge application porting effort begins to adapt to the new distributed MPP architecture
  - Large injection of funding to commercial companies, their trajectory simply would not have met our goals without it

Accelerated Strategic Computing Initiative (ASCI) Program Plan https://www.osti.gov/servlets/purl/768266

1993 Intel paragon (Acoma)

- While the ASCI program is being established, Sandia fully commits to the distributed parallel computing
- Sandia's first #1 on Top 500 in 1994 at 143.4 GigaFLOPS
  - Remains in top 5 until 1997
  - 2nd Gordon-Bell prize
- 1890 Compute nodes
  - Intel i860 RISC microprocessors
- OSF-1 planned as the OS but was inefficient in practice
  - SUNMOS to the rescue (<256kbytes of memory) maximizes memory available for applications
  - Sandia continued light-weight Kernel development and testing of PUMA on the Paragon which eventually replaced SUNMOS on ASCI Red as Cougar (not at all confusing)
- Portals version 0 – network hardware programming interface
- Diskless booting
  - Majority of nodes, compute nodes
  - Specialized functionality of nodes, compute, IO, service
- Red/Black switching
  - Dedicated nodes on both restricted and classified network
  - Center switchable shared portion
  - Repeating theme for Sandia systems through Red Storm

Introduction of beanie babies

Nelson Mendela and F.W. de Klerk joint Nobel Prize for peaceful end to apartheid in South Africa

Harley Davidson turns 90

Intel Paragon, Sandia Labs #1 Top 500

1996 ASCI RED (Janus)

- ASCI Red is deployed at Sandia National Laboratories
  - First deployment of the color themed ASCI program platforms
- First teraflop/s computer, taking the No.1 spot on the 9th TOP500 list in June 1997 with a Linpack performance of 1.068 teraflop/s
  - Originally 200 MHZ Pentium processors
  - Chip upgrade (Pentium II Xeon @333 MHZ) increased to 3.1 teraflop/s
  - 104 cabinets, 76 compute
  - Over 1600 sq feet of space
  - parallel filesystem delivered 1GB/sec, unprecedented at the time
    - Disk failures were problematic
  - 400MB/sec interconnect
- Lets call this the first production MPP system at Sandia
- Part of Building 880 Annex was refurbished to hold ASCI Red
  - Huge facility effort
- Puma was adopted by Intel and deployed as Cougar
  - Ported Linux later in life, didn't perform as well as Cougar
- Portals version 2 (what happened to 1?)
- Retained #1 on TOP500 list from June 1997 to June 2000
- Replaced as #1 by another ASCI color schemed platform, IBM's ASCI White at Lawrence Livermore National Laboratory, November 2000
- ASCI Red was retired from service in September 2005, after having been on 17 TOP500 lists over eight years
  - Another characteristic of many systems at Sandia is longevity spurred on by in place upgrades
- 1998 Meritorious Sandia Achievement Award

President Bill Clinton
First signatory of comprehensive Nuclear Test Ban Treaty
Has yet to be ratified by senate

Down closes the year over 6k
1k point gain in one year.
Over 34k now

NINTENDO 64

Nintendo 64 release

Cougar

portals

ASC Red, first machine to exceed **1 TFLOPS**

1997 Cplant™

- Yes the name was actually trademarked
- As a result of the great success of ASCI Red, Sandia was cut out of the Tri-lab platform rotation
  - Can you tell this still annoys me?
- No worries we will build our own out of commodity parts
- Multiple instantiations: Hawaii, Alaska, Siberia and Antarctica
  - Antarctica – 1U servers, very dense for the time, Alpha 21264 processors
  - Myrinet interconnect
- Emulate ASCI Red environment including using much of the software that was developed
- Sandia developed most of the software but used Linux (RedHat) as the OS
  - RedHat only about 3 years old at this time, not developed for HPC
- These were the early/beginning days of clusters almost everything was roll-your-own
- Cplant™ was the first Terascale cluster computer
- Top500 (listed as Self Made)
  - Nov 1998 #97 54 GFLOP
  - June 1999 #129 @54 GFLOP (unofficial #53 @124 GFLOP)
  - Nov 1999 #44 @ 232 GFLOP
  - June 2001 #41 @ 512 GFLOP
  - Nov 2001 #30 @ 706 GFLOP
  - Nov 2002 #48 @ 996.9 GFLOP (1800 nodes)
- Cplant™ consistently pushed the bleeding edge of scale
- All Cplant™ clusters of note were Dec/Compaq Alpha based
- Sandia truly acted as the platform Integrator
- Antarctica actually had 4 heads, open, restricted, classified and development
  - Switching between these heads was very patchwork
- All nodes booted from a single disk on the administration node
- Possibly the most important aspect of our Cplant™ effort was preparing staff for Red Storm

Hawaii

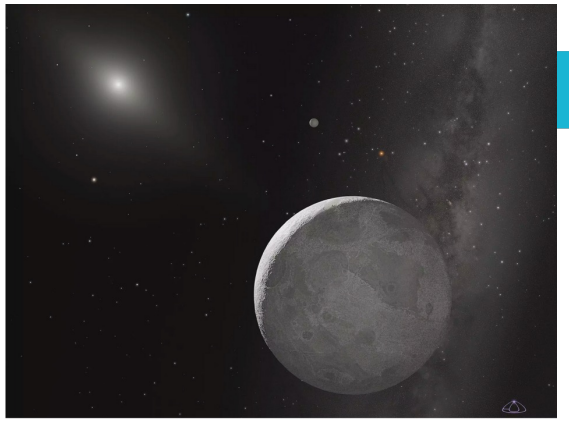
Alaska

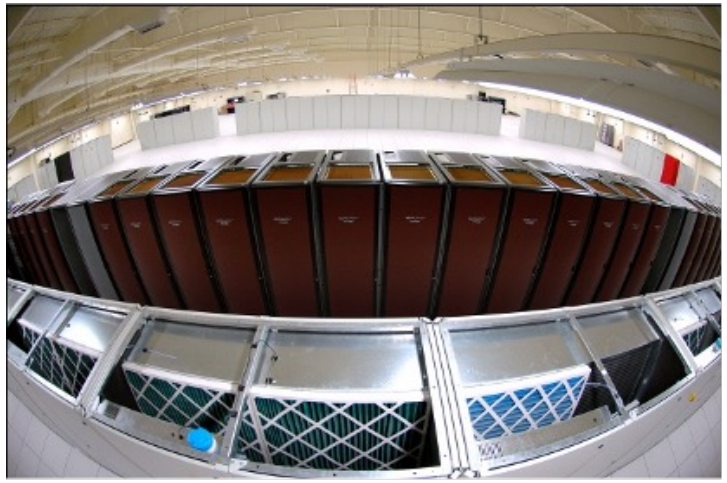
Siberia

2005 Red storm

- 2001 - Requirements definition
- 2002 – Cray awarded contract
- Marriage made, well maybe not in heaven
  - Cray was on the rocks
  - Sandia was really sticking their neck out
    - Size of the project larger than the net worth of the company (Cray)
  - Both parties were highly motivated to succeed
- Off the shelf processor and memory
  - AMD Opteron - First 64 bit x86 processor
- Custom interconnect
  - Seastar (PowerPC processor)
  - Custom network became an ongoing theme for Cray value add
- 10,880 single core processors
  - 10,368 compute, 512 service and IO
  - Multiple upgrades in place, dual and quad-core
  - Technically a heterogeneous platform
- Top500 (each major upgrade)
  - November 2005: Rank 6 (36.19 TFLOPS)
  - November 2006: Rank 2 (101.4 TFLOPS)
  - November 2008: Rank 9 (204.2 TFLOPS)
- Cray contracted Sandia to deliver much of the software
- Cougar is now Catamount
  - Compute node operating system
  - Followed by Catamount N-Way when processors are upgraded to multi-core
  - All compute nodes booted in just a few minutes
- Portals version 3
- Sandia even implemented the SeaStar driver
- Early hardware deliveries booted with Cplant™ management software
  - Cluster Integration Toolkit (CIT)

Eris, 10th planet discovered





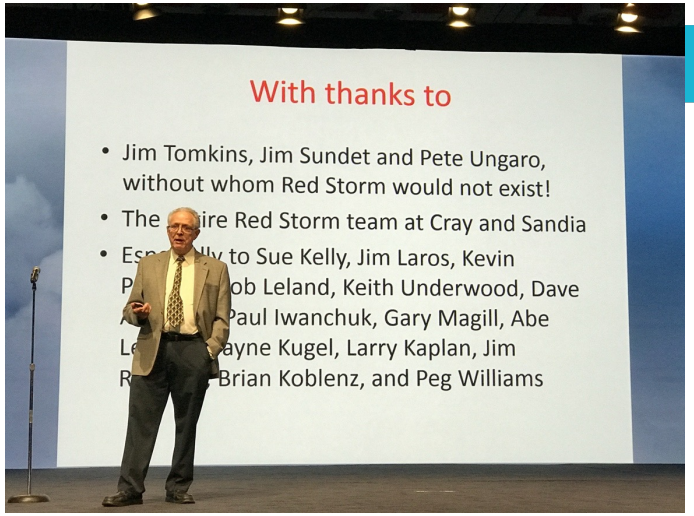
Red Storm in Building 725 West

Catamount

portals

Red storm (cont.)

- Bill Camp – “fastest development cycle of any supercomputer”
  - Less than 2 years, typically 5
  - Recipient of 2016 IEEE Computer Society Seymour Cray Computer Engineering Award “for visionary leadership of the Red Storm project, and for decades of leadership of the HPC community.”
- Jim Tomkins - “I had this nickname at Cray, ‘the devil incarnate,’ because I was something of a hard nose,”
  - I’m sure Cray thought this was something of an understatement, personally he is a sweet guy
- Peter Ungaro CEO Cray - “Without Red Storm I wouldn’t be here in front of you today. Virtually everything we do at Cray — each of our three business units — comes from Red Storm. It spawned a company around it, a historic company struggling as to where we would go next. Literally, this program saved Cray.”
- Among the machine’s technical achievements was the operation in 2008 known as Burnt Frost, in which Red Storm programmed a 152-inch rocket to shoot down an errant satellite traveling at 17,000 miles per hour, 153 miles above the earth.
- The result: after the successful take-down with no collateral damage, a military commander exulted, “We can hit a spot on a bullet with a bullet.”
- Cray: As of 2012 over a billion dollars in sales
- Most successful Supercomputer line EVER
- Cray (now HPE) prime on LLNL and ORNL, integrator for Argonne, for DOE exascale systems
- On a personal note, I had the honor of powering Red Storm down for the final time

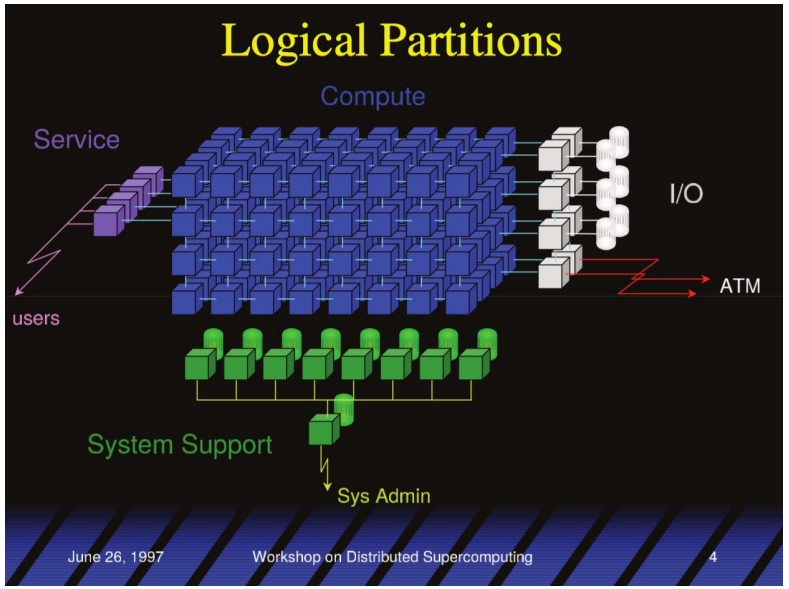


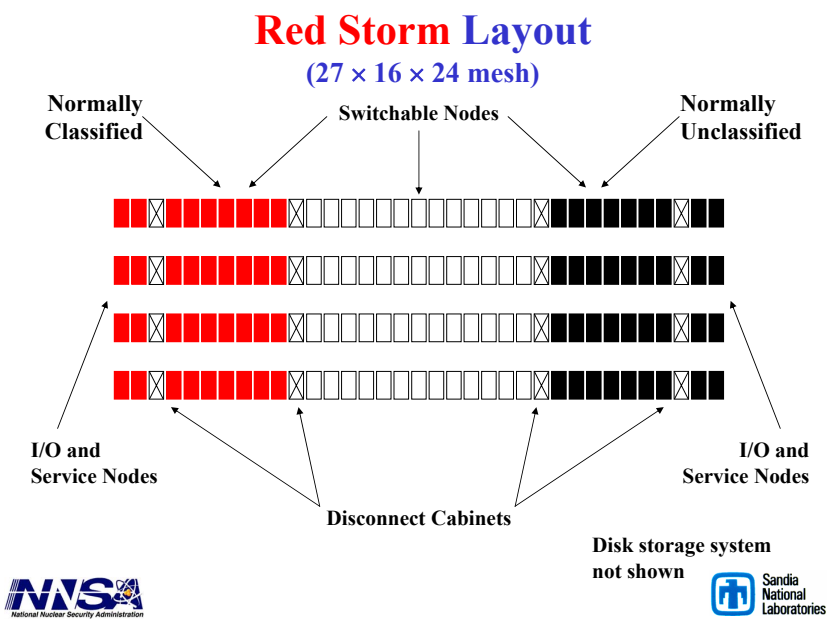
2009 R&D 100 Award for Catamount N-Way Light Weight Kernel
2006 NOVA Award Red Storm Design and Development Team
2006 Sandia Meritorious Achievement Award, Red Storm Design, Development and Deployment Team
…. to name a few

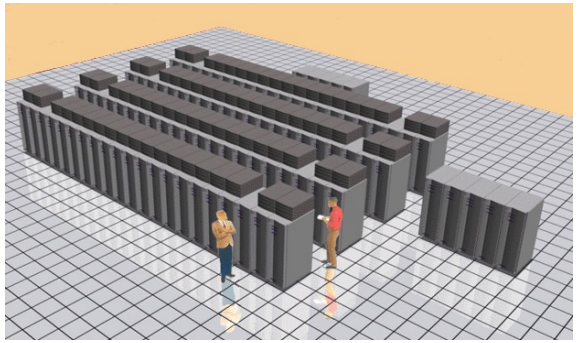
A quick Aside

- One of the most important aspects of large platform efforts like this at Sandia, in my opinion, are the new efforts that are spawned either at of necessity or from new research that is enabled
  - LWK and Portals efforts were integral parts of deploying these platforms
- The characteristics of the Catamount LWK enabled many systems software investigations at Sandia
- High Performance Computing - Power Application Programming Interface Specification
  - Catamount allowed researchers to exploit emerging processor capabilities to experiment with early attempts at conserving power/energy on HPC systems
  - Essentially put the OS to sleep when an application was not resident on the node
  - Led to additional experimentation and savings when dual and quad core processors were deployed
  - This effort was enabled in no small part by our ability to modify the RAS system and actually capture real power data
  - This early work led to Sandia having great impact in an emerging field
  - Power API spec remains the only published API for Power/Energy measurement and control
  - Awards:
    - 2018 R&D 100 Award
    - 2018 R&D 100 Award, special recognition
    - 2011 NNSA Environmental Stewardship Award
- This is one of many examples
- I would claim some of Sandia's most impactful research was enabled by Red Storm and platforms like it

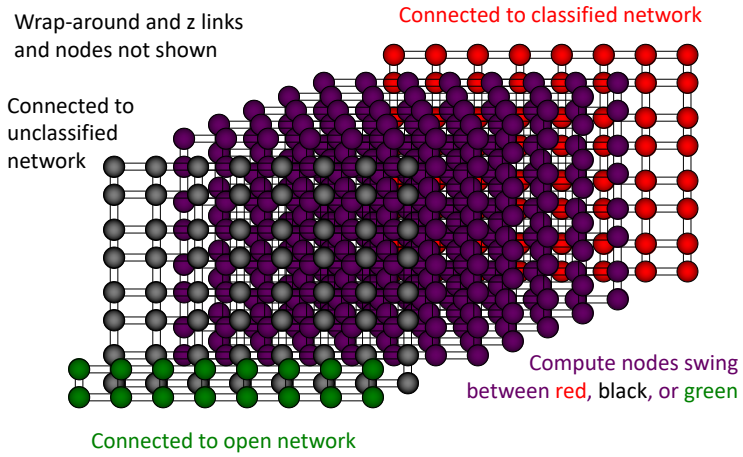
Common architectural themes







Zermatt (Cplant 2000)

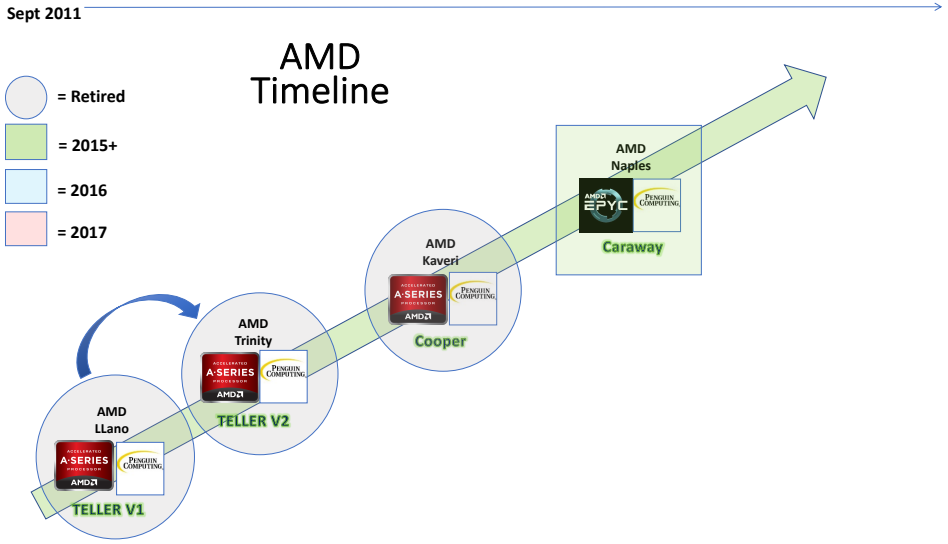


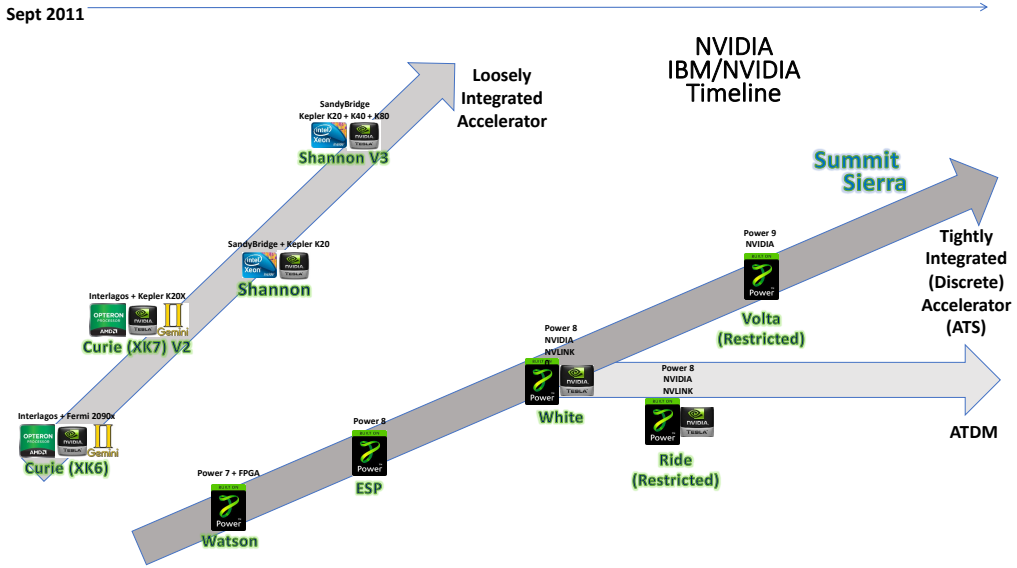
2010 – So now what?

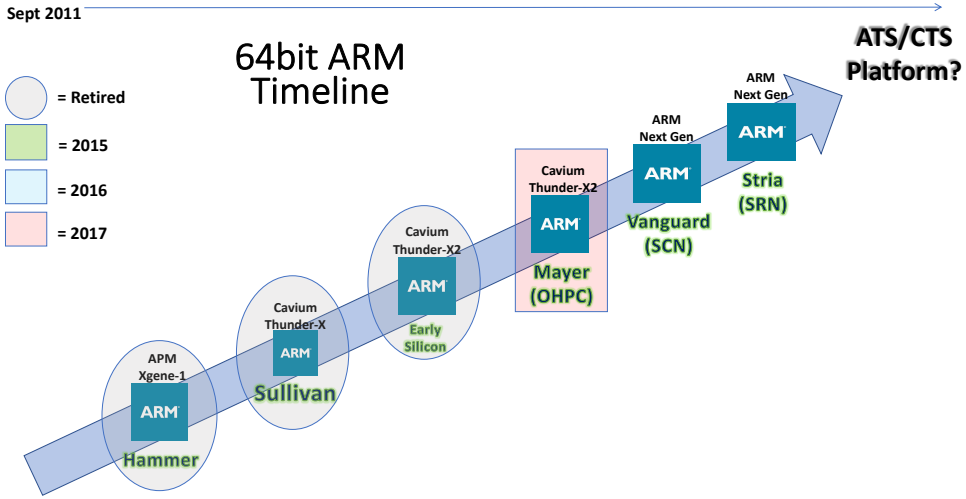
Swine Flu Pandemic, actually started in 2009

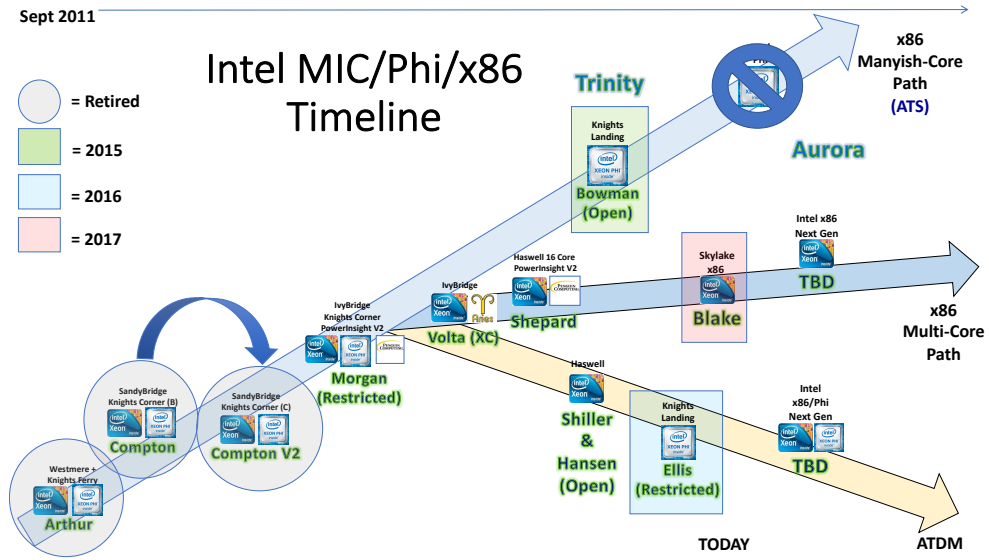
- Clusters are now commodity
- Red Storm has gone dark
  - Serving a different mission but still running until 2012
- A diversity in node architectures emerging
  - Somewhat similar to today (AI/ML)
- In the absence of large scale deployments how can we have an impact?
- Advanced Architecture Test Bed Program is born
- Node to Rack scale investigations into technologies that have potential to impact our mission
- VERY often pre-general availability silicon, truly bleeding edge
- Tight collaborations with technology providers and integrators
- Leverage Mini-apps to provide feedback to technology provider partners
- Essentially, Sandia found a way to continue tight partnerships and influence technology without fielding large-scale platforms
- … to be a scout for future computer architectures

Test bed evolution





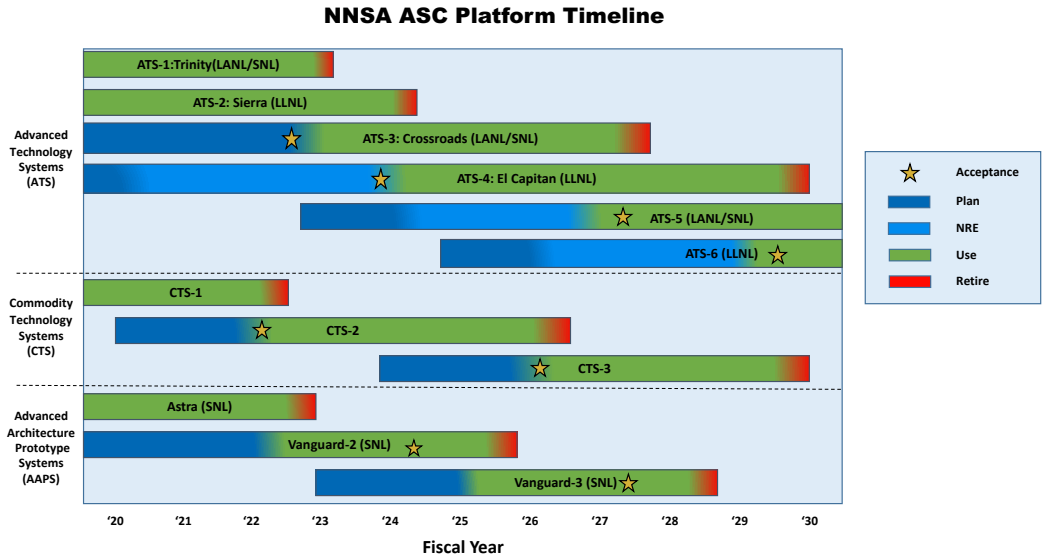




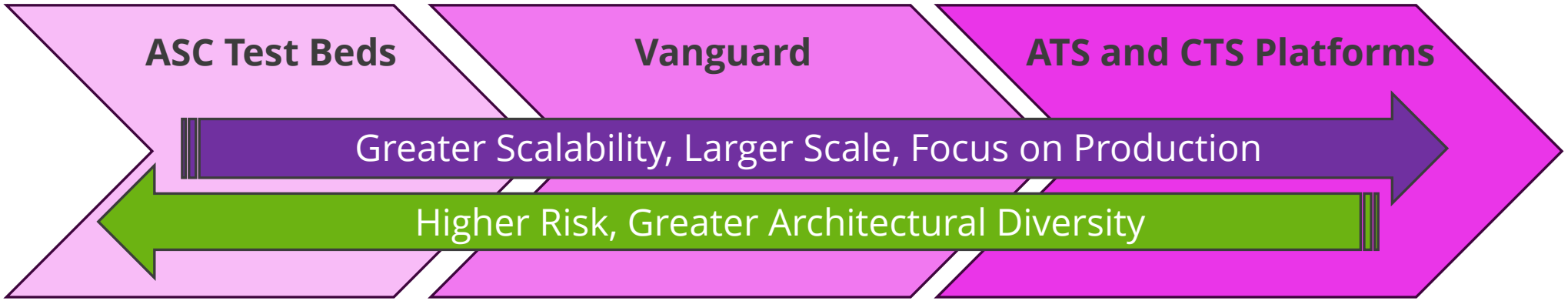
2018 Vanguard/Astra

Super Blue Moon January 2018, first time since 1866

- Sandia chosen to host the first Arm-based supercomputer
  - The return of large-scale platforms at Sandia
- Huge amount of pressure and a compressed timeline
  - Feels like Red Storm all over again
- Advanced Architecture Prototype Program is born as a result of Sandia's success
  - Goal: prove the viability of emerging advanced architecture technologies for the NNSA mission
- 725 West hosted first x86_64 based Supercomputer – Red Storm
- 725 East will host first Arm64 based Supercomputer – Astra
  - Maybe 725 North can host the first RISCV based Supercomputer and you can be part of it?

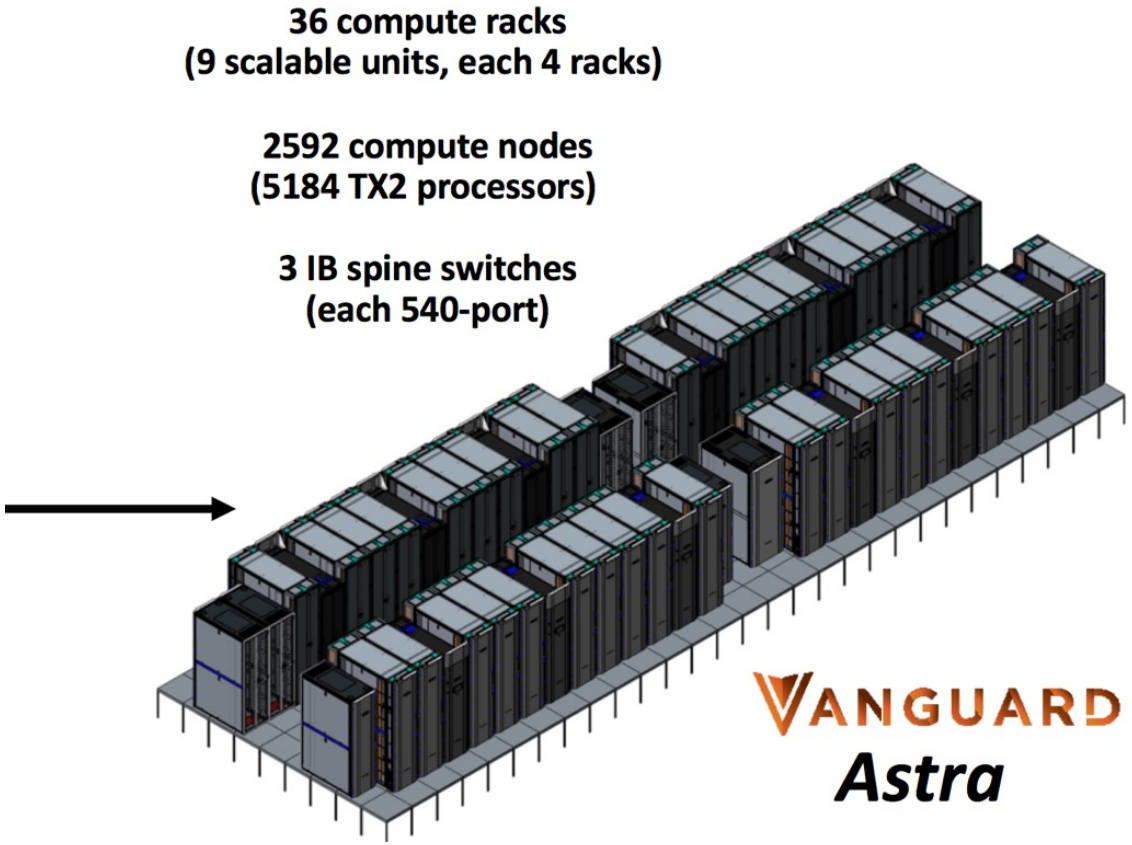


February 2021



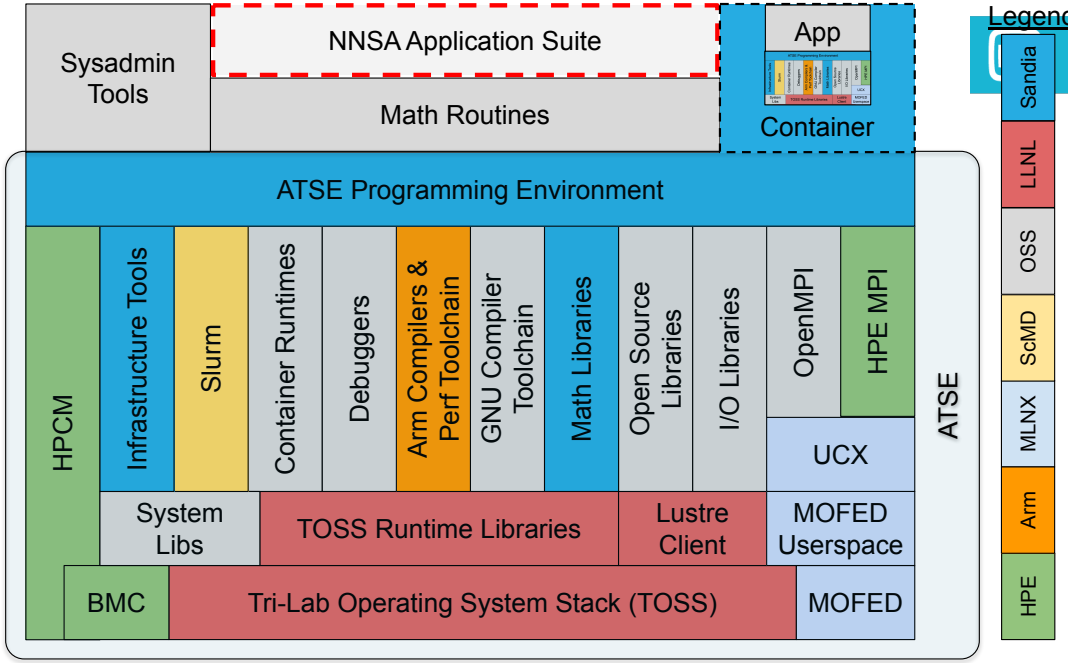
Astra at a glance

- **2,592** HPE Apollo 70 compute nodes
  - Cavium Thunder-X2 **Arm** SoC, 28 core, 2.0 GHz
  - 5,184 CPUs, 145,152 cores, 2.3 PFLOPs system peak
  - 128GB DDR Memory per node **(8 memory channels per socket)**
  - Aggregate capacity: 332 TB, Aggregate Bandwidth: 885 TB/s
- Mellanox IB EDR, ConnectX-5
- HPE Apollo 4520 All–flash storage, Lustre parallel file-system
  - Capacity: 990 TB (usable)
  - Bandwidth 244 GB/s



Lets not forget the software

- Advanced Tri-lab Software Environment (ATSE)
- Question: what do you do when your supercomputer shows up without a software stack?
- Answer: If you have Sandia's history, you roll your own
- Simply put ATSE enabled the many successes we achieved with Vanguard/Astra
- Less than a month from the time Astra hit the floor to at-scale Top500 run
- Not possible without a **TEAM** effort
- First tera-scale Arm platform (2.3 TFLOPs peak)
  - Astra debuted at #204, 1.529 TFLOPs, on Top500 list, November 2018
  - Improved to #156, 1.758 TFLOPs, June 2019
- Astra continues to provide production cycles
- Awards:
  - 2021 Employee Recognition Award, Astra support of ASC L1 Milestone
  - 2019 Employee Recognition Award, Astra Supercomputer Team
  - 2018 Defense Programs Award of Excellence, Exceptional Achievement



Take away

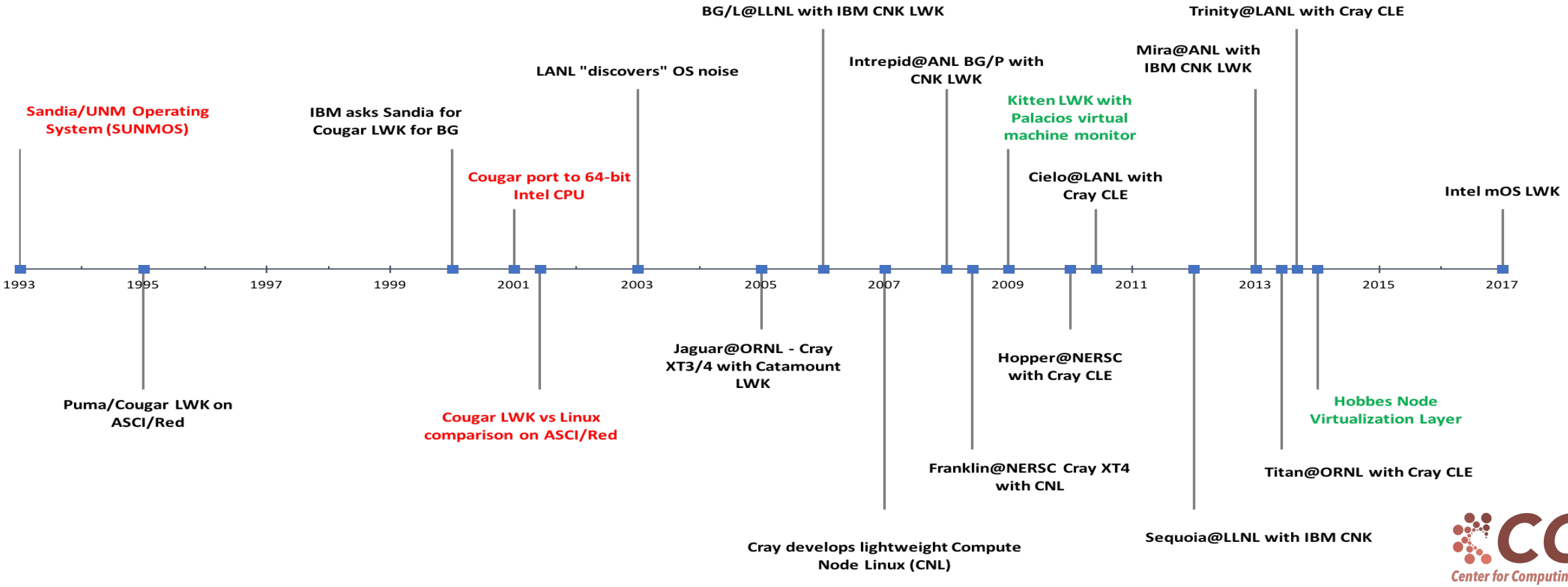
- Sandia is very much an engineering laboratory
- 1400 didn’t just buy computers we have architected them and contributed to the field in both hardware and software innovation
- 9300 has been our partner though most if not all of these endeavors
- 9300 also fielded an important cutting edge platform - Red Sky
  - Time and fear of misrepresenting it prevented me from covering it here today
- Portals impact outside Sandia includes Atos, Cray, Intel and Mellanox
- LWK development influenced IBM Blue Gene LWK and current Intel mOS (Rolf Riesen former Sandia staff)
- PowerAPI is now a community standard
- General Architecture influence, clusters and tightly integrated platforms
  - Mentioned our influence on Cray, now HPE architectures
  - Fugaku has many architectural similarities, runs a light-weight kernel
- I challenge you to find another DOE laboratory that has so significantly partnered, and contributed, with technology providers and integrators
- My hope is that you all continue our tradition of impact in the years to come on programs like Vanguard and what follows

Questions?

Sandia's LWK Approach Has Had Broad Impact

Sandia is the only DOE laboratory to partner with vendors to deploy a custom OS in production

- SUNMOS LWK on Intel Paragon; Cougar LWK on ASCI/Red; Catamount on Cray Red Storm
- Other vendors have followed the LWK model: IBM CNK for BG/{L,P,Q}; Cray's Linux Environment
- Every large-scale DOE distributed memory machine in the past 25 years has deployed a lightweight OS



CCR Center for Computing Research

Significant Vendor Impact of Sandia's Portals Networking Technology

All of these production vendor-supported systems used Portals as the network hardware programming interface.
Portals enabled the first TeraFLOPS platform (ASCI Red) and the first non-accelerated PetaFLOPS platform (Jaguar).

Intel Paragon

Portals 0

Intel ASCI Red

Portals 2

Cray Red Storm

Portals 3

Cray XT3, XT4, XT5

Portals 3

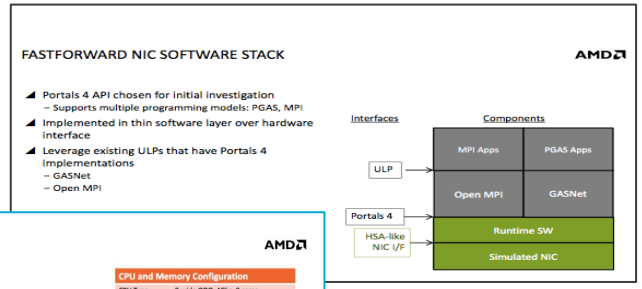
Atos Tera1000

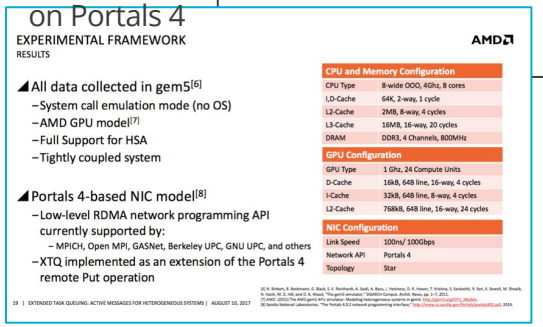
Portals 4

portals

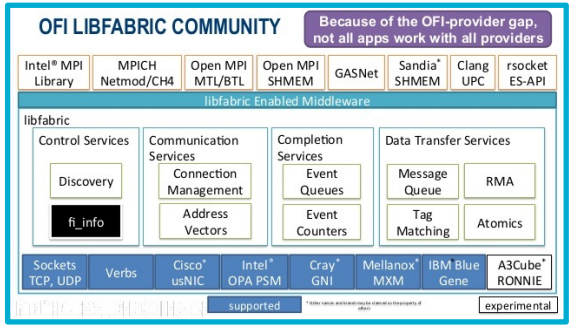
Unlike other low-level network programming interfaces, Portals is intended to enable co-design rather than serve as a portability layer.
The influence and impact of Portals can be seen in vendor co-design activities, other low-level network programming interfaces, and emerging network hardware.

AMD FastForward Project based on Portals 4

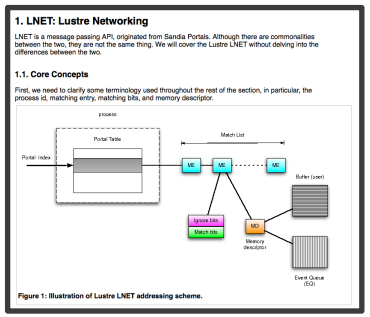


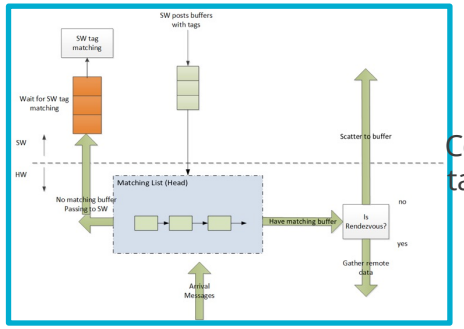


OFI Libfrabric API based on Portals 4



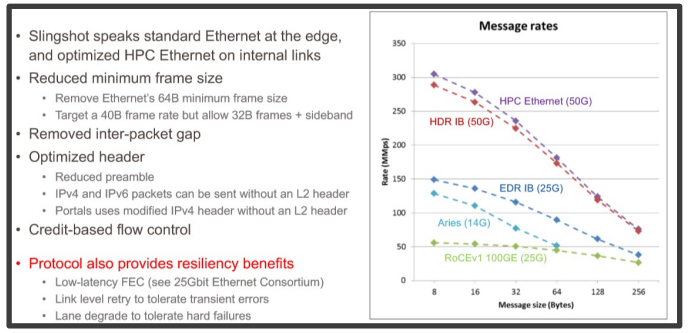
Lustre File System network based on Portals 4



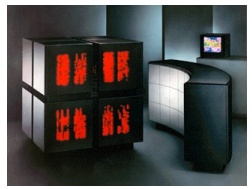


Mellanox ConnectX-5 MPI tag matching in hardware

Cray Slingshot Supports Portals 4 header



Systems Software History



CM-2
1989

nCUBE-2
1990

iPSC-860
1992

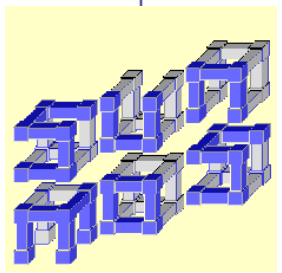
Paragon
1993

ASCI Red
1996

Cplant
1998

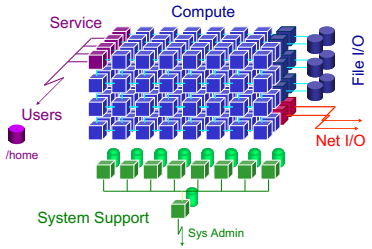
Red Storm
2005

1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007

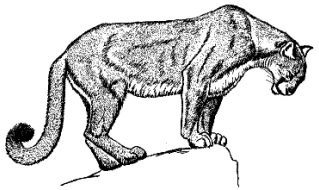


SUNMOS
1991 - 1997

Portals
1992 –



Partition Model
1993 –



Puma, Cougar,
Catamount
1993 –

Cplant
1997 – 2005

Kitten
2007 –