SAND2017-8695PE

Exceptional service in the national interest

Sandia National Laboratories

HPC Co-Design

Ron Brightwell, R&D Manager

CCR Center for Computing Research

U.S. DEPARTMENT OF ENERGY | National Nuclear Security Administration

Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

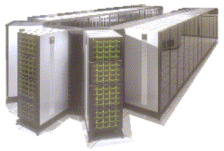
Track Record of HPC Success

Sandia National Laboratories


CM-2


nCUBE-2


iPSC-860


Paragon


ASCI Red


Cplant


Red Storm

1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009

Gordon Bell Prize

Patent
Meshing

R&D 100
Dense Solvers

Gordon Bell Prize

R&D 100
Xyce

R&D 100
Parallel Software

R&D 100
Storage

World Record
Teraflops

R&D 100
CPU Allocator

Gordon Bell Prize

Mannheim
SuParCup

Patent
Data Mining

R&D 100
Signal Processing

SC96 Gold Medal
Networking

Karp Challenge

World Record
281 GFlops

R&D 100
Salvo

R&D 100
Catamount LWK

R&D 100
Meshing

Patent
Partitioning

R&D 100
Trillinos

Patent
Parallel Software

World Record
143 GFlops

R&D 100
Aztec

Fernbach
Award

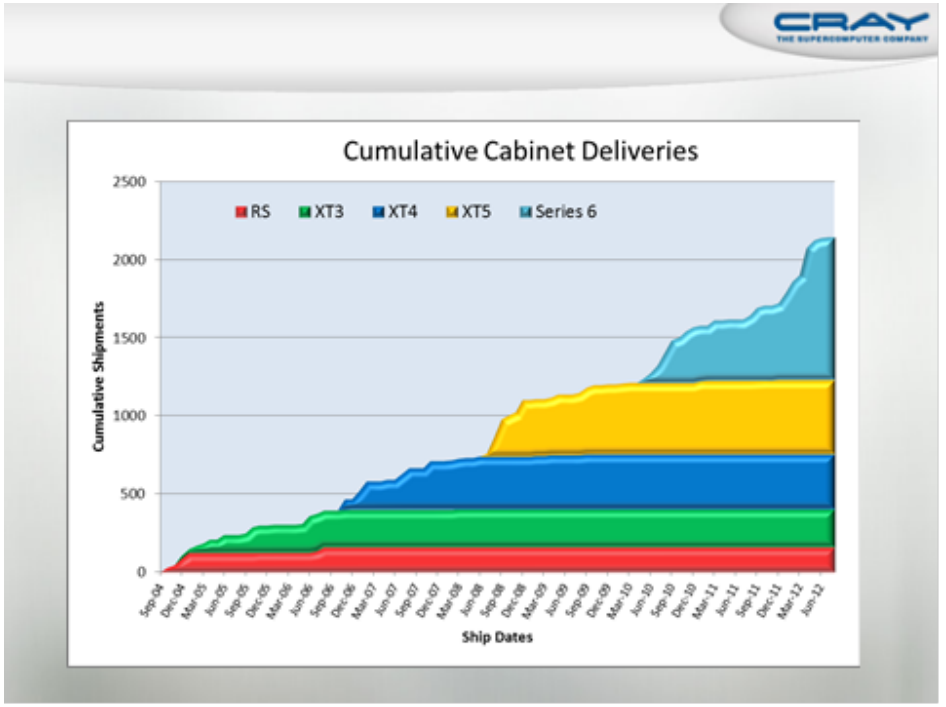
Patent
Paving

CCR
Center for Computing Research

Red Storm – Prototype for Cray XT Series

Sandia National Laboratories

- Architected by Sandia, engineered jointly with Cray
  - Sandia contributed to the design of the SeaStar network interface and router
- Sandia also developed
  - Lightweight kernel compute node OS
  - Scalable parallel job launching system
  - Portals high-performance interconnect programming interface
  - SeaStar firmware
- 140+ systems to 80 different customers worldwide
  - Including ORNL, NERSC, and LANL
- Following Red Storm, Cray's market share rose from 6% in 2002 to 21% in 2007*
- Revenue of $1B +
- Basis of Cray's business today



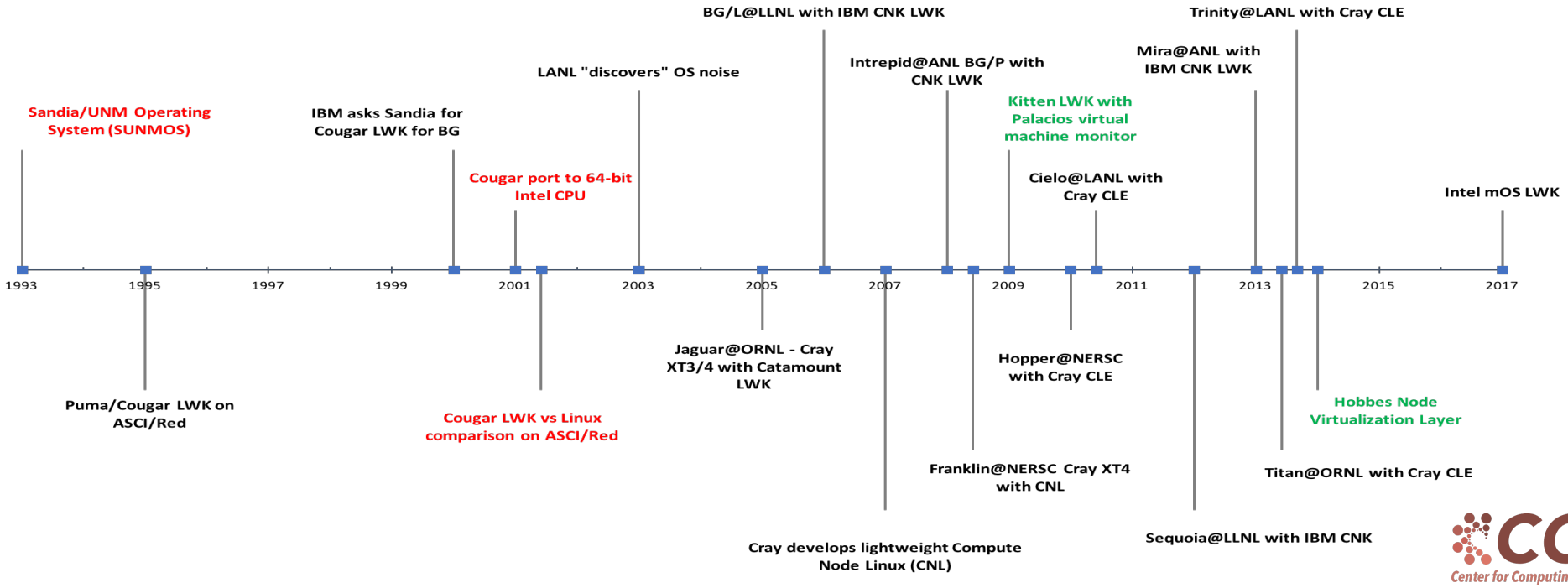
*Source: IDC #209251 *Technical Computing Systems: Competitive Analysis*, November 2007

CCR Center for Computing Research

Sandia's LWK Approach Has Had Broad Impact

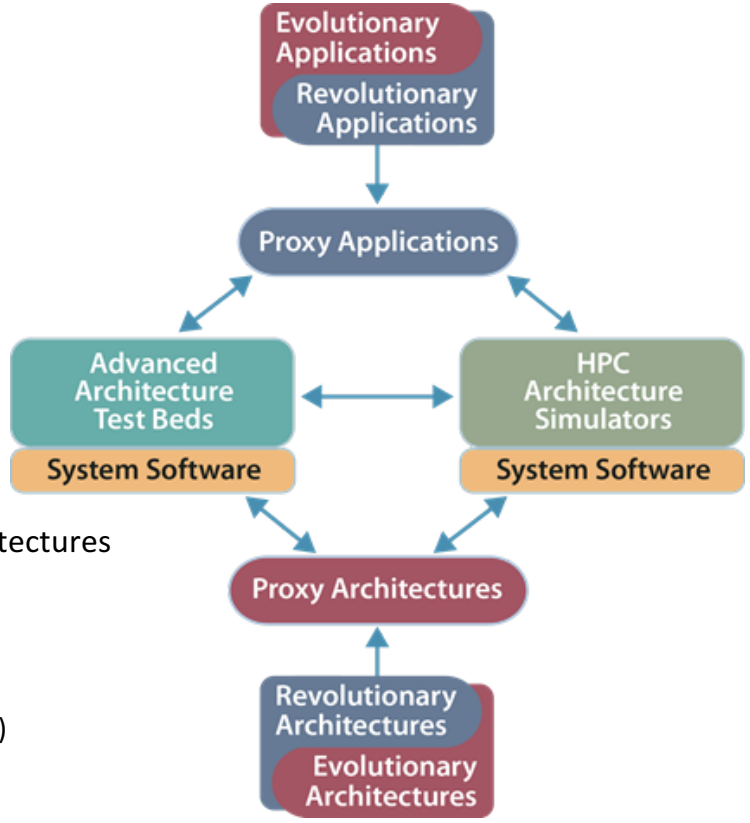
Sandia is the only DOE laboratory to partner with vendors to deploy a custom OS in production

- 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



Sandia HPC Co-design Capabilities

- Proxy Applications (Mantevo):
  - Application source for architecture-centric optimization and analysis
  - http://mantevo.org
- HPC Architectural Simulation Framework:
  - Structural Simulation Toolkit (SST)
  - Flexible to accommodate fidelity/speed tradeoffs
  - http://SST-simulator.org
- ASC Advanced Architecture Testbeds:
  - Evolving examples of COTS "state-of-the-art"
  - http://www.sandia.gov/asc/computational_systems/HAAPS.html
- Abstract Machine Model (AMM) Definitions and associated Proxy Architectures
  - Developed by SC/ASCR Computer Architecture Lab (SNL & LBL)
  - http://crd.lbl.gov/assets/pubs_presos/CALAbstractMachineModelsv1.1.pdf
- System software components
  - Kitten lightweight operating system (https://github.com/HobbesOSR/kitten)
  - Qthreads user-level thread library (https://github.com/Qthreads/qthreads)
  - Portals interconnect API (https://github.com/Portals4/portals4)



Portals Interconnect Programming Interface

Sandia National Laboratories

- Developed by Sandia, U. New Mexico, Intel
- Previous generations of Portals deployed on several production massively parallel systems
  - 1993: 1800-node Intel Paragon (SUNMOS)
  - 1997: 10,000-node Intel ASCI Red (Puma/Cougar)
  - 1999: 1800-node Cplant cluster (Linux)
  - 2005: 10,000-node Cray Sandia Red Storm (Catamount)
  - 2009: 18,688-node Cray XT5 – ORNL Jaguar (Linux)
- Focused on providing
  - Lightweight "connectionless" model for massively parallel systems
  - Low latency, high bandwidth
  - Independent progress
  - Overlap of computation and communication
  - Scalable buffering semantics
  - Protocol building blocks to support higher-level application protocols and libraries and system services
- At least three hardware implementations currently under development
- Portals influence can be seen in IB Verbs (Mellanox) and libfabric (Intel & others)

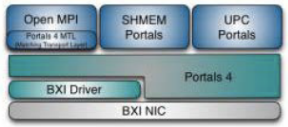
CCR Center for Computing Research

Portals Hardware and Co-Design Activities

Sandia National Laboratories

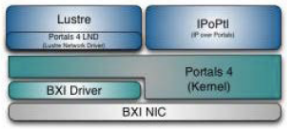
BXI application environment

BXI comes with a complete software stack to provide optimal performance and reliability to all traditional HPC components.



BXI Computing stack

- Parallel applications can take full advantage of the capabilities of the BXI network using MPI, SHMEM or UPC communication libraries.
- All components are implemented directly using the Portals 4 API.
- Kernel services are also implemented using the kernel Portals 4 implementation.
- A Portals 4 LND (Lustre Network Driver) provides the Lustre parallel filesystem with a direct / native access to Portals 4.
- The IPoPtl (IP over Portals) component makes it possible to have large scale, efficient and robust IP communication for legacy software.



BXI Kernel services

For more information: Please contact hpc@atos.net

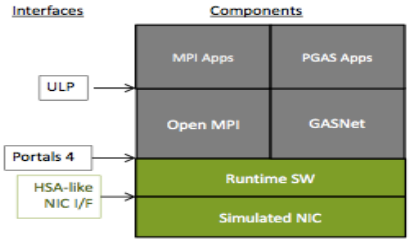
Your business technologists. **Powering progress**

Atos

FASTFORWARD NIC SOFTWARE STACK

AMD

- Portals 4 API chosen for initial investigation
  - Supports multiple programming models: PGAS, MPI
- Implemented in thin software layer over hardware interface
- Leverage existing ULPs that have Portals 4 implementations
  - GASNet
  - Open MPI



EXPERIMENTAL FRAMEWORK

RESULTS

AMD

- All data collected in gem5[6]
  - System call emulation mode (no OS)
  - AMD GPU model[7]
  - Full Support for HSA
  - Tightly coupled system
- Portals 4-based NIC model[8]
  - Low-level RDMA network programming API currently supported by:
    - MPICH, Open MPI, GASNet, Berkeley UPC, GNU UPC, and others
  - XTQ implemented as an extension of the Portals 4 remote Put operation

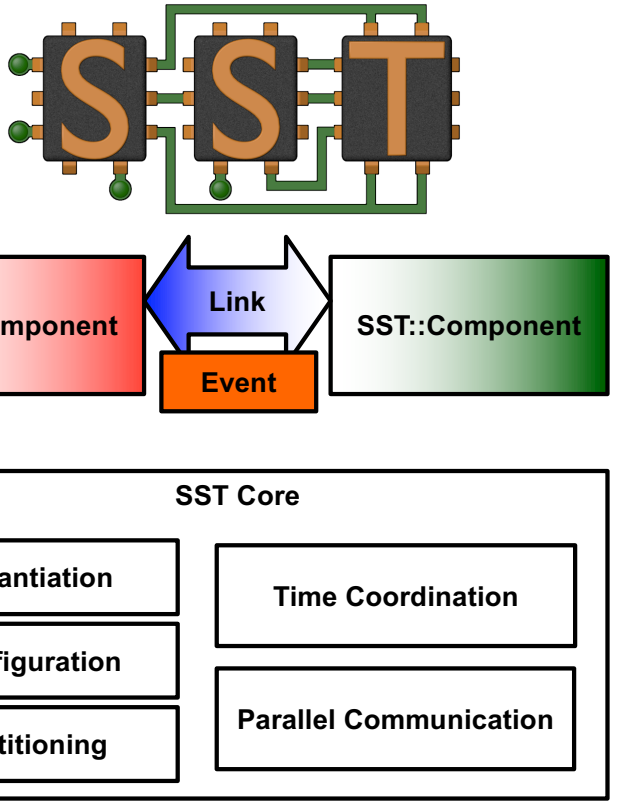
CPU and Memory Configuration	
CPU Type	8-wide OOO, 4Ghz, 8 cores
I,D-Cache	64K, 2-way, 1 cycle
L2-Cache	2MB, 8-way, 4 cycles
L3-Cache	16MB, 16-way, 20 cycles
DRAM	DDR3, 4 Channels, 800MHz
GPU Configuration	
GPU Type	1 Ghz, 24 Compute Units
D-Cache	16kB, 64B line, 16-way, 4 cycles
I-Cache	32kB, 64B line, 8-way, 4 cycles
L2-Cache	768kB, 64B line, 16-way, 24 cycles
NIC Configuration	
Link Speed	100ns/ 100Gbps
Network API	Portals 4
Topology	Star

[6] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The gem5 simulator," SIGARCH Comput. Archit. News, pp. 1–7, 2011.
[7] AMD. (2015) The AMD gem5 APU simulator: Modeling heterogeneous systems in gem5. http://gem5.org/GPU_Models.
[8] Sandia National Laboratories, "The Portals 4.0.2 network programming interface," http://www.cs.sandia.gov/Portals/portals402.pdf, 2014.

CCR Center for Computing Research

SST in a Nutshell

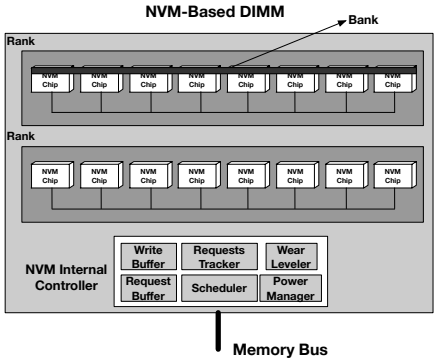
- Use Supercomputers to Design Supercomputers
- Parallel Discrete-Event Simulator Framework
  - Flexible framework allows multitude of custom simulators
  - Demonstrated scaling to over 512 processors
- Comes with many built-in simulation models
  - Processors, Memory, Network
- Open API
  - Easily extensible with new models
  - Modular framework
  - (Non-Viral) Open-source core
- Time-scale independent core
  - Handles Micro-, Meso-, Macro-scale simulations
- "Best of Breed" – Bring together work from Labs, Industry, Academia



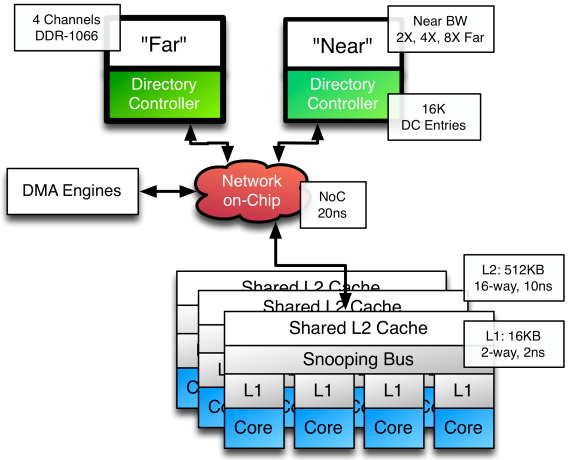
SST Use Cases

Sandia National Laboratories

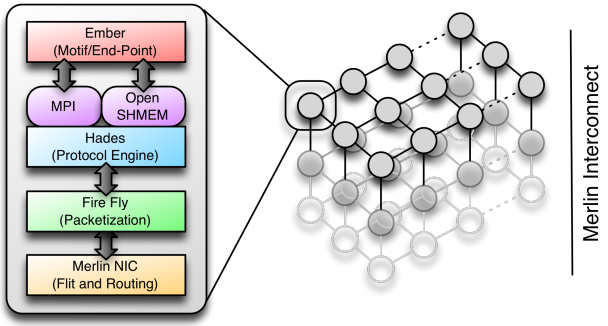
Emerging NV Memory Technologies



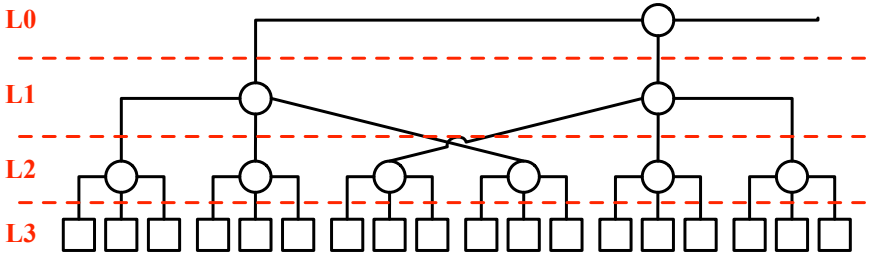
Multi-Level Memory (HBM+DDR+NV)



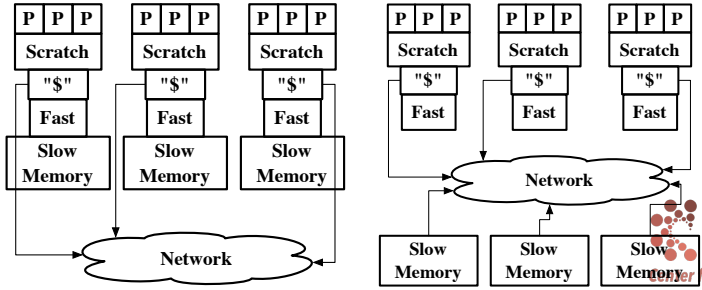
Communication/ Interconnect Modeling



Photonic Network Topology & Routing



Disaggregated Memory



CCR Center for Computing Research

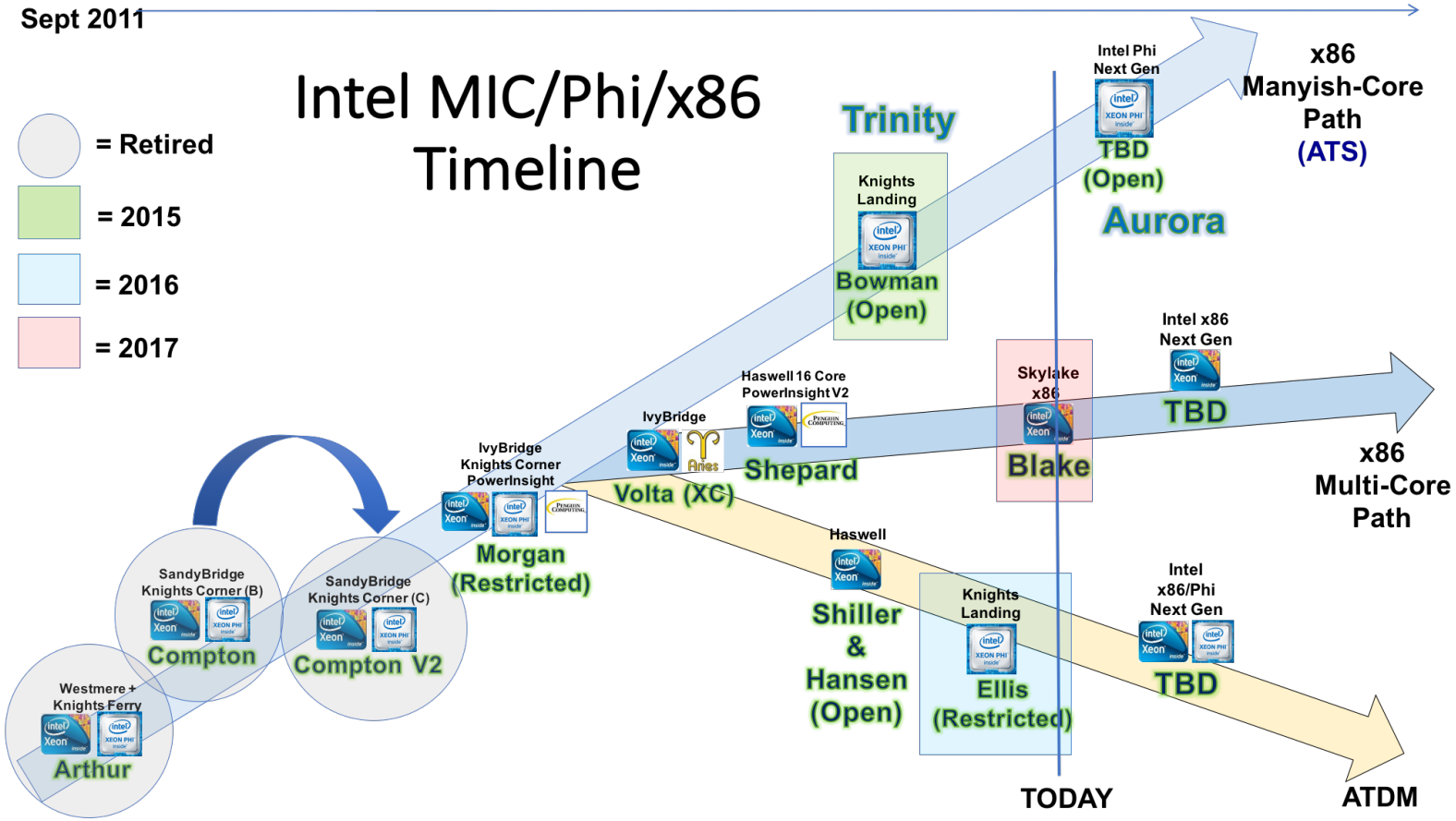
Advanced Architecture Testbeds

Sandia National Laboratories

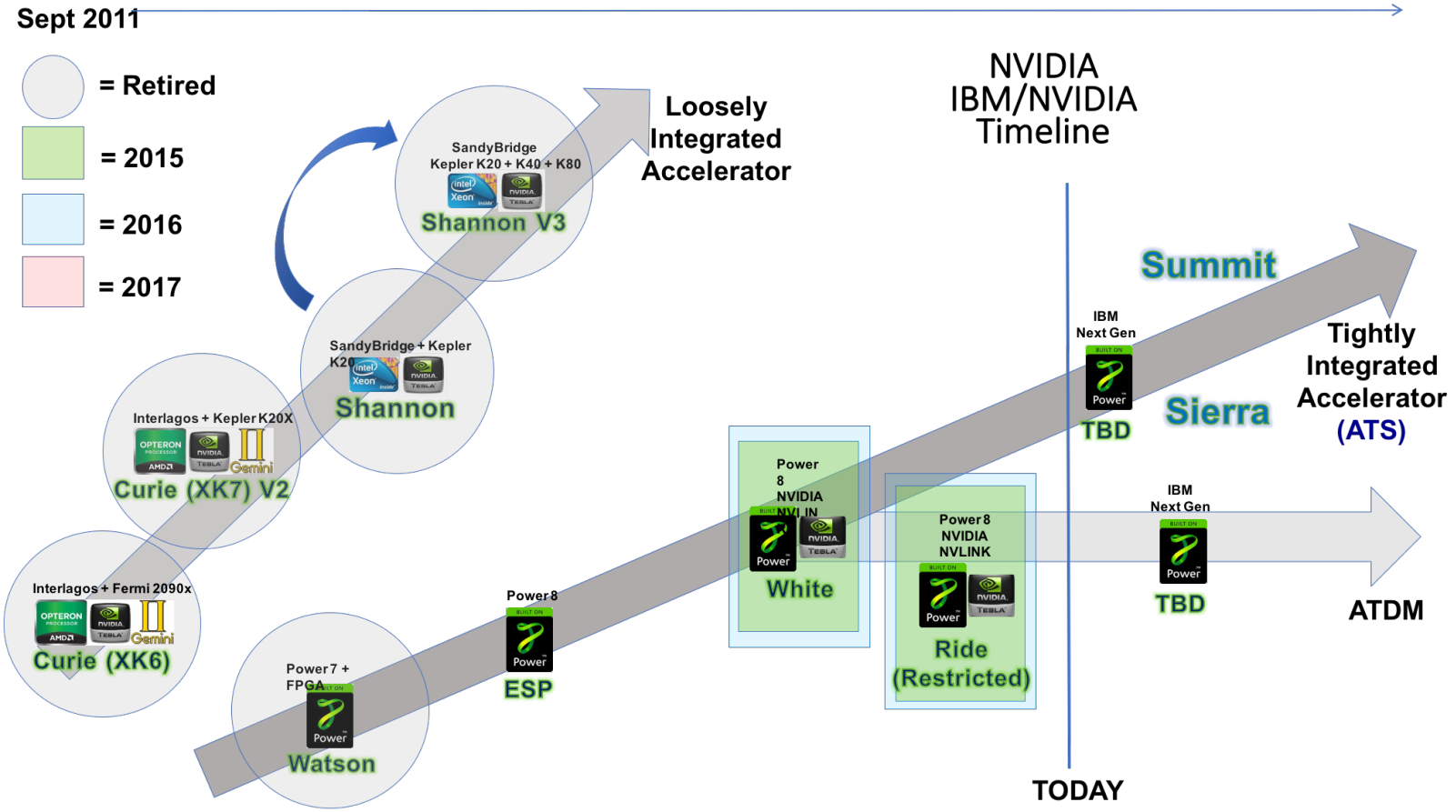
- Pre-production vendor systems
  - Intentionally closer to prototypes than production
  - Not intended for production application cycles
  - Priority is to explore wide and diverse set of emerging architectural alternatives
  - Rack-scale
- Facilitates co-design
  - Can potentially identify enhancements or influence configurations of follow-on components
- Helps prioritize vendor efforts
- Reduces application porting time on production systems

CCR Center for Computing Research

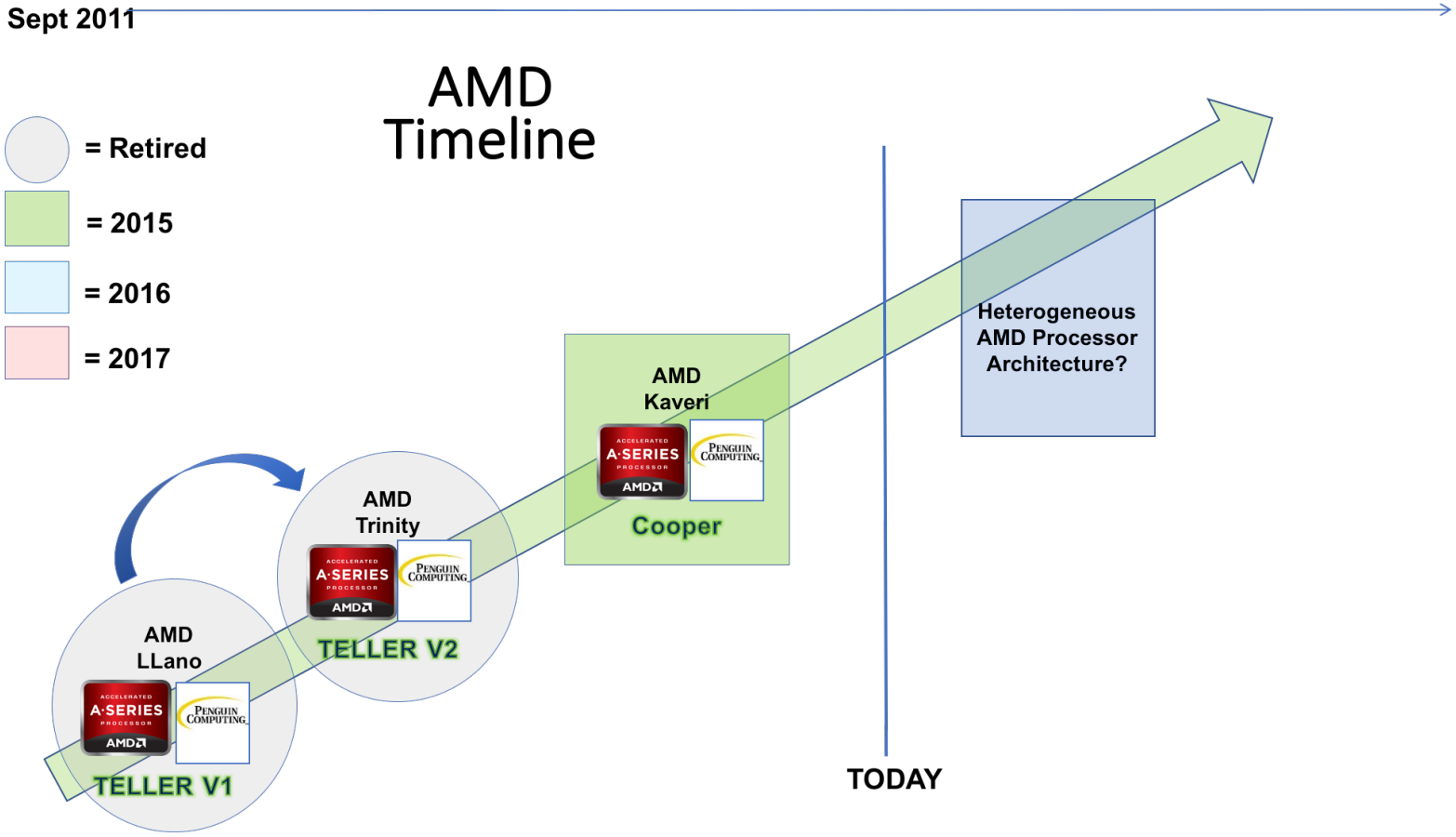
Intel Testbed Timeline



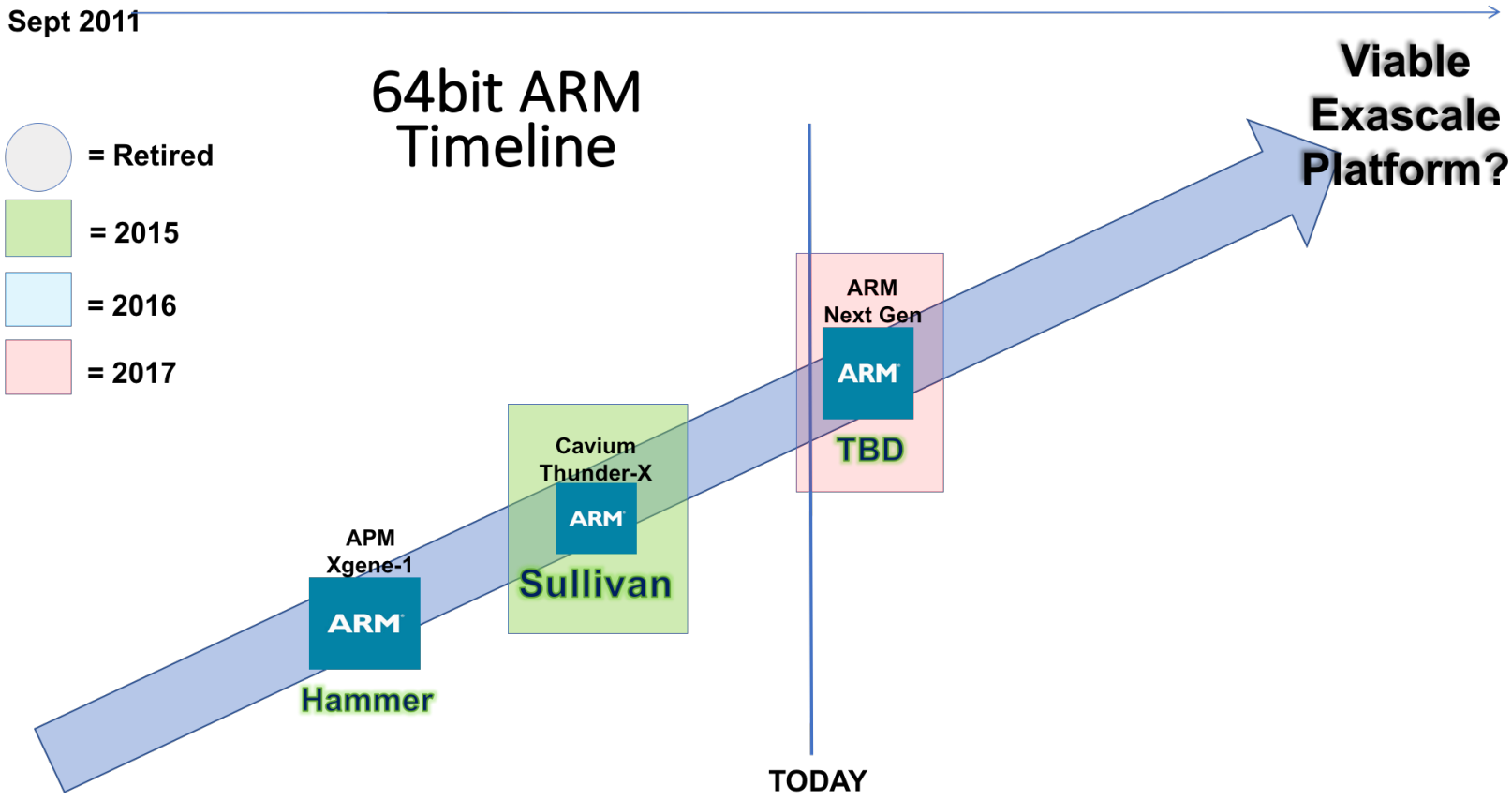
IBM Testbed Timeline



AMD Testbed Timeline



ARM Testbed Timeline



Large-Scale ARM Platform at Sandia

Sandia National Laboratories

- 10-15 PFLOPS
- Key Requirements for this 2019 Platform
  - Proof of Concept for future leadership class DOE platform
  - Competitive HPC 64 Bit ARM processor technology
  - Integrated On-package Memory
  - First of a kind, Advanced Interconnect technology
  - Large focus on maturing the ARM software ecosystem
- Opportunities to *prototype* innovative test hardware
  - Logic in NIC to improve interconnection network performance electrical or optical
  - Logic in/near memory to support sparse linear algebra acceleration

CCR Center for Computing Research