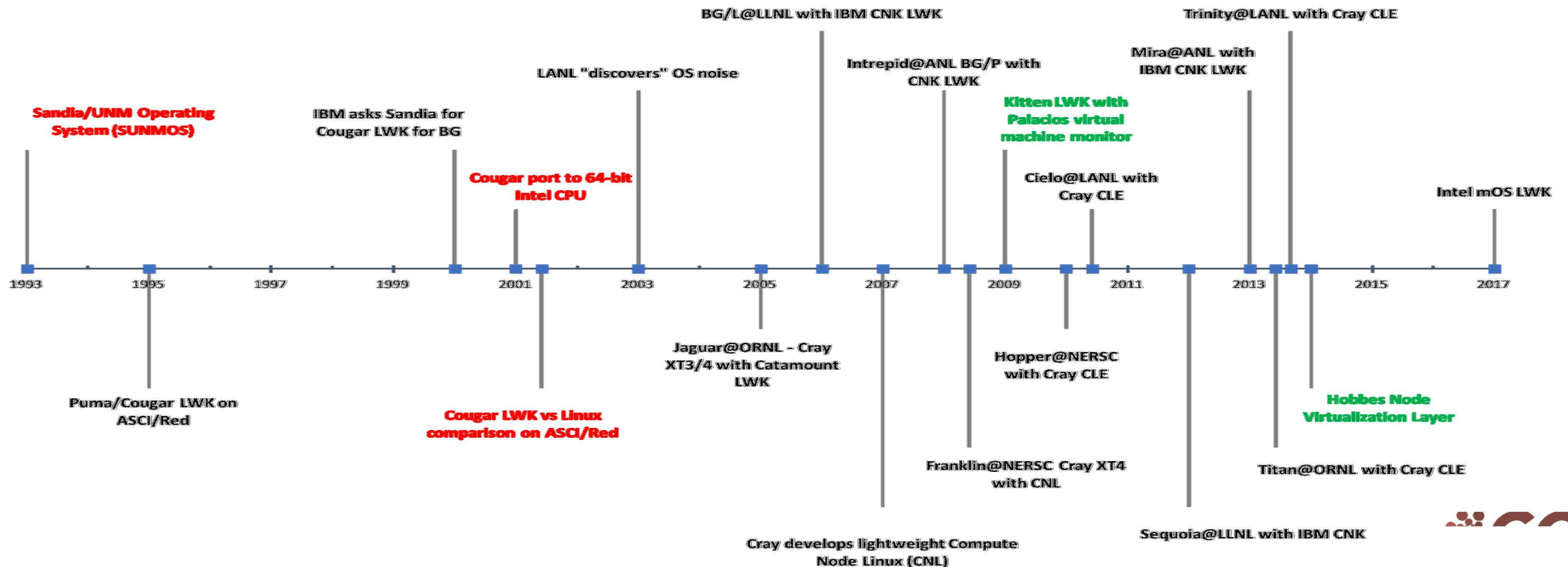


# Sandia's LWK Approach Has Had Broad Impact

SAND2019-10078PE

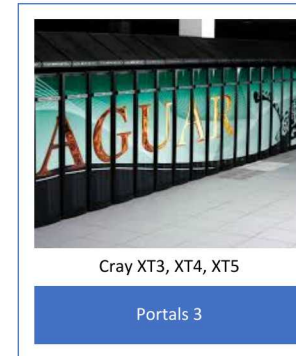
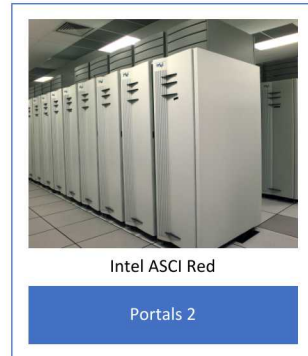
Sandia is the only DOE laboratory to partner with vendors to deploy a custom OS in production

- 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



# Significant Vendor Impact of Sandia's Portals Networking Technology

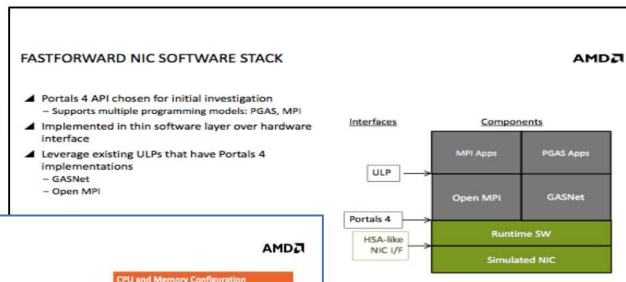
All of these production vendor-supported systems used Portals as the network hardware programming interface. Portals enabled the first TeraFLOPS platform (ASCI Red) and the first non-accelerated PetaFLOPS platform (Jaguar).



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



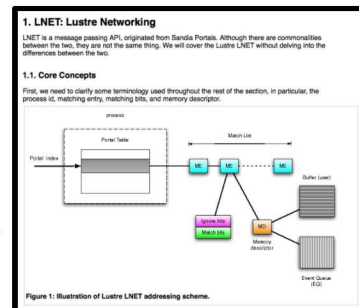
## EXPERIMENTAL FRAMEWORK RESULTS

- All data collected in gem5<sup>[6]</sup>
  - System call emulation mode (no OS)
  - AMD GPU model<sup>[7]</sup>
  - Full Support for HSA
  - Tightly coupled system

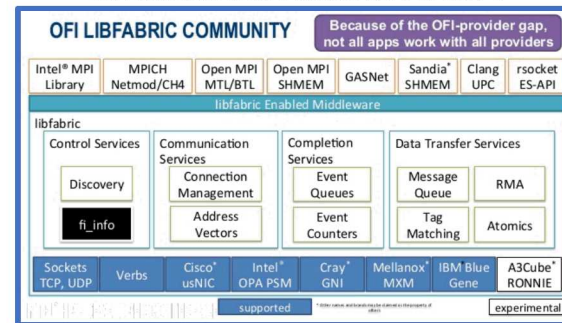
## Portals 4-based NIC model<sup>[8]</sup>

- 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

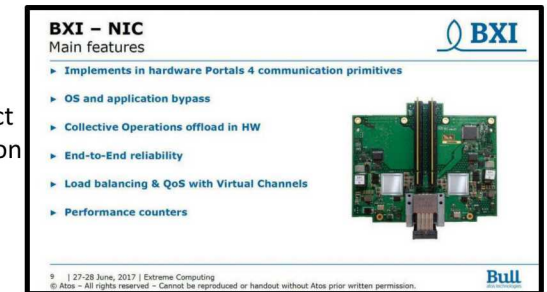
CPU and Memory Configuration	
CPU Type	8-wide OOO, 40Hz, 8 cores
L2-Cache	64K, 2-way, 1 cycle
L3-Cache	2MB, 8-way, 4 cycles
L2-Cache	16MB, 16-way, 20 cycles
DRAM	20GB, 4 Channels, 800MHz
GPU Configuration	
GPU Type	1 GHz, 24 Compute Units
D-Cache	16KB, 64B line, 16-way, 4 cycles
L2-Cache	32KB, 64B line, 8-way, 4 cycles
L2-Cache	768KB, 64B line, 16-way, 24 cycles
NIC Configuration	
Link Speed	100Gb/100Gbps
Network API	Portals 4
Topology	Star



## OFI Libfabric API based on Portals 4

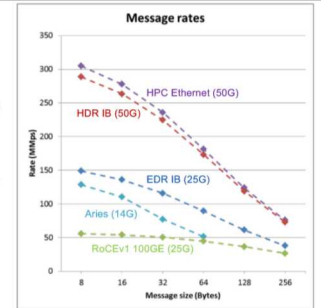


Atos Bull eXascale Interconnect (BXI) based on Portals 4



Cray Slingshot Supports Portals 4 header

- Slingshot speaks standard Ethernet at the edge, and optimized HPC Ethernet on internal links
- Reduced minimum frame size
  - Remove Ethernet's 64B minimum frame size
  - Target a 40B frame rate but allow 32B frames + sideband
- Removed inter-packet gap
- Optimized header
  - Reduced preamble
  - IPv4 and IPv6 packets can be sent without an L2 header
  - Portals uses modified IPv4 header without an L2 header
- Credit-based flow control
- Protocol also provides resiliency benefits
  - Low-latency FEC (see 25Gbit Ethernet Consortium)
  - Link level retry to tolerate transient errors
  - Lane degrade to tolerate hard failures



Lustre File System network based on Portals 4

Mellanox ConnectX-5 MPI tag matching in hardware