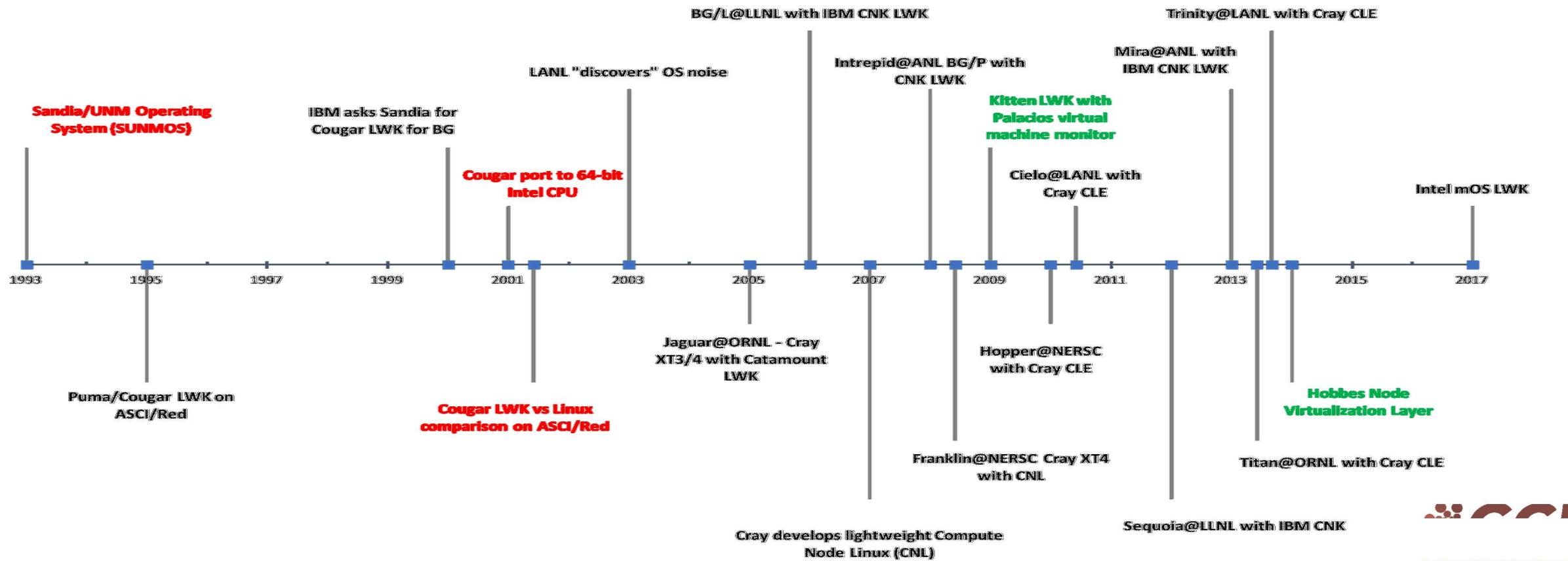


Sandia's LWK Approach Has Had Broad Impact

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).



Intel Paragon

Portals 0



Intel ASCI Red

Portals 2



Cray Red Storm

Portals 3



Cray XT3, XT4, XT5

Portals 3



Atos Tera1000

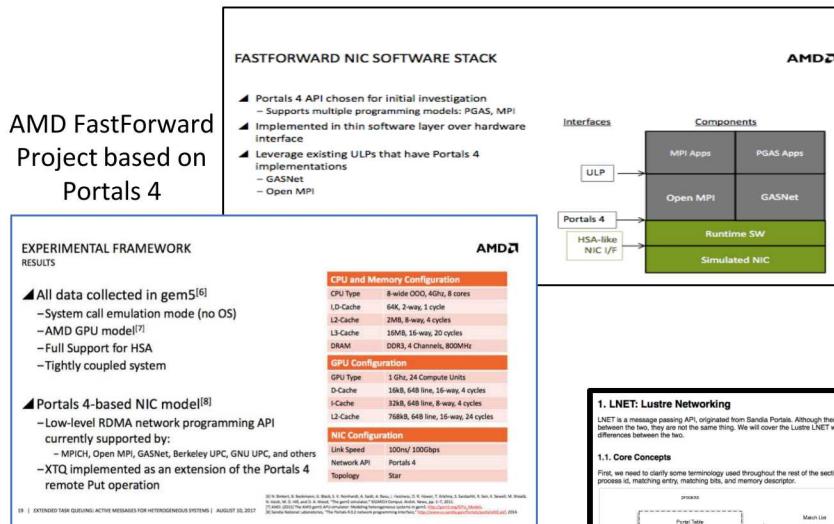
Portals 4



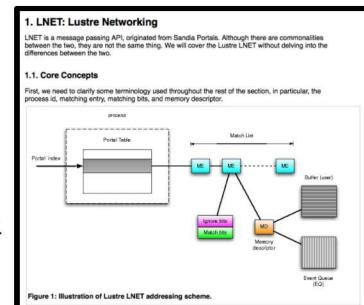
Unlike other low-level network programming interfaces, Portals is intended to enable co-design rather than serve as a portability layer.

The influence and impact of Portals can be seen in vendor co-design activities, other low-level network programming interfaces, and emerging network hardware.

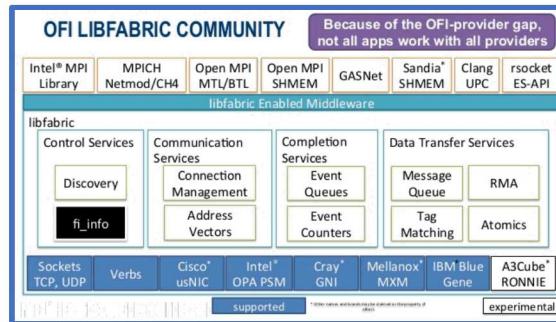
AMD FastForward Project based on Portals 4



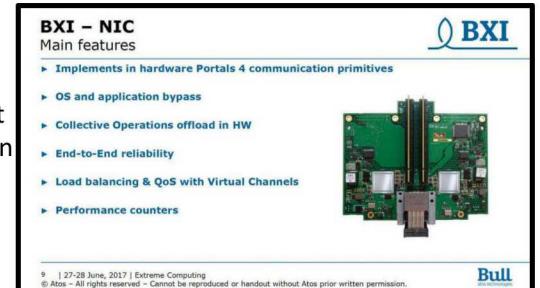
Lustre File System network based on Portals 4



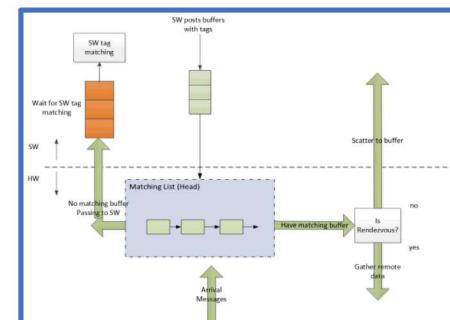
OFI Libfrabric API based on Portals 4



Atos Bull eXascale Interconnect (BXI) based on Portals 4



Cray Slingshot Supports Portals 4 header



Mellanox ConnectX-5 MPI tag matching in hardware

- 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 25Gb Ethernet Consortium)
 - Link level retry to tolerate transient errors
 - Lane degrade to tolerate hard failures

