



Accelerators and Networks: How should they interact?

CAC 2007 Panel Presentation

Keith D. Underwood



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.





Accelerators FOR Communications

- **Offload is good!**
- **Basically, all of MPI should be on the NIC**
 - **Except maybe derived datatype processing...**
- **Another core on the processor is not the answer**
 - **What are the big challenges facing processors?**
 - **Off-chip latency no longer decreasing**
 - **Off-chip bandwidth not growing as fast as core count**
 - **Processing should be close to the hardware it is driving**
 - **Latency tolerance of I/O instructions is basically non-existent**
 - **Both polling and interrupts can be expensive**
 - **Processors are not very good at realistic MPI scenarios**

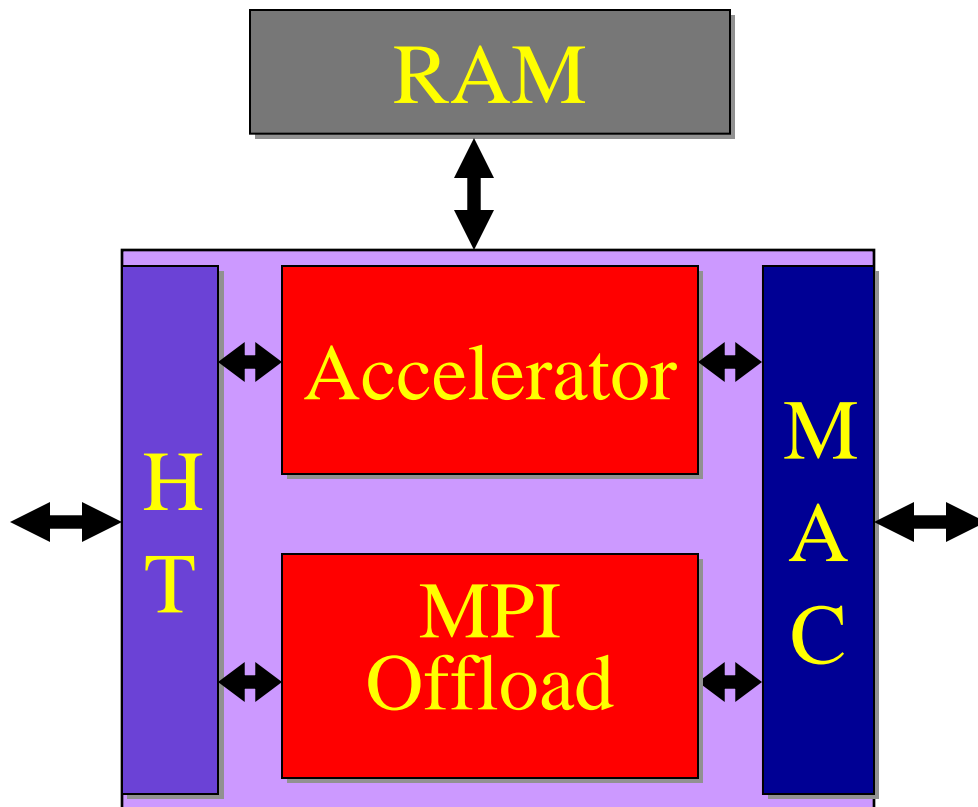


What makes an “Accelerator”?

- **Good at something the processor is not**
 - Higher peak FLOPs, or
 - Unstructured data, or
 - Streaming data, etc.
- **In that sense, “offload” is an “accelerator”, but more common examples of accelerators include:**
 - FPGAs
 - Clearspeed
 - Cell
- **Generally, something that actually does application processing**



Acceleration Coupled with Communications



- Accelerator has access to host and network
 - Does not need to communicate with MPI
 - Can do application specific thing in an application defined way
 - Could also couple to MPI
- Particularly amenable to streaming processing
 - Process the data as it goes through



Example Applications

- **Parallel FFT (actually, matrix transpose)**
 - Contiguous data needs to be shuffled and transmitted
 - Received data needs to be interleaved and placed in contiguous location
- **Parallel sparse matrix operations**
 - Each iteration requires the gathering of small data items from many other nodes
 - Data is result of previous iteration
 - Numerous iterations needed
- **Swap-and-add**
 - Boundary items are transferred to neighbor and accumulated



It can be Built

- **Explored in academia**
 - **Intrusion detection – FPGAs or IXP (various)**
 - **Programmable routers – FPGA based (John Lockwood)**
 - **Application specific network services – IXP (Karsten Schwan)**
 - **Intelligent NICs – FPGA based (Underwood)**
- **Necessary hardware has already been shipped in products, but seldom enabled**
 - **Cray (OctigaBay) XD1 placed an FPGA next to the NIC**
 - **SRC has a separate fabric for the FPGAs**
- **Major issues in the past**
 - **Lack of vendor support (in the form of libraries)**
 - **Programming environment for application developers**
 - **Unnecessary cost**



Where will we be in 5 years?

- **Someone, somewhere will finally get TCP offload right**
 - But, the \$5 NIC chip in your PC still won't do it
 - Which means half of the clusters out there won't have it
- **Collective offload will begin to appear**
 - But vendors will resist offloading floating-point collectives for no obvious reason
- **Read-modify-write support will become more common**
 - But it won't be used, because MPI will still dominate the application base