

IPDPS / CAC 2003 Panel Session

*Mitchell Gusat
IBM Research*

Zurich Research Lab

Why no fast & cheap cluster comms yet?
Wanted: Mainstream NPs and Programmable Switches

- 1. HPC libs and stacks *not natively adopted* [yet] by NPs
 - ❖ ASIC ecosystem => cost & perf. based on economy of scale

When a **mainstream NP** for SANs & StANs ?

- 2. Stupid fabrics
 - ❖ Whether E or O, the ICTN features are frozen by design
 - Few, if any, switch-router features are programmable
 - e.g., routing tables

When an ISA for Switch/Routers to expose the ICTN to the app?

Do all HPC apps need the same QoS, FC, fairness, buf. mgnt., RD, flow (de)coupling etc...?

Ex., Graphical Processing Unit (or a DSP):

-transparent for most apps, but programmable for graphic or audio libs (DirectX)

Affordable & versatile HP ICTN: iff the COM convergence...*

- The Golden Age of ICTNs (books and companies mushrooming), yet..
 - ❖ At their 50th anniversary this year, ICTNs are between art and primitive science (routers today are akin to CPUs in early '80s)
- 21st century: the next confluence? **comp + comm => COM***

Why?

- ❖ Dozens of sticky standards (and still coming...)
 - Each standard: unstable specs, owing to unanticipated functional demands
- ❖ Per market & application, the list of requirements increasingly exceeds the capabilities of ASIC systems (also valid for optical systems).
 - "I no longer understand switching..." - reputable switching expert from Stanford
- Wanted: *One [programmable] switch/NP* to cover multiple apps and build h/w and s/w ecosystems (like Pentium, AMD and PowerPC)
- Open: What's the DirectX of HPC?

Improvements - Evolutionary

- 1. Improved arbitration, allocation, scheduling, fairness
 - ... growing issues: **power**, packaging
 - ❖ Unexpectedly open (juicy research topics)
 - Load balancing: DF-AR in gen. topos - space (local, global, mixed - blurry horiz.), time (hist., current, predictive) state info
 - CC: pro- / re- / active, e2e FC, lossless, drop (impact on Rel. Delivery?)
 - LL-FC: CP seen as the poor relative of DP => re-balance the conveyer belt w/ HP FC
 - Scheduling: interleaving of prio and deadline-based channels
 - Fairness: local vs. global, partitioned clusters, (de)coupled flows, novel FQ-ing
 - Contention resolution of O and E switches: arbitration, allocation, centralized, distributed
 - **Interplay of the above!**
- 2. H/w ICTN support for global synchros, weak consistency/OOO-tolerance, object-oriented coherency (var. grain), mcast & QoS
- 3. Latency-hiding: MT-ICTNs => multithreaded FC and scheduling

Improvements - Breakthroughs

- 1. Programmable ICTNs
 - ❖ ISA for Switching-Routers and NPs
 - ❖ Expose the ICTN = NP + Switch-Router
- 2. ICTN-aware Proc. and Mem.
 - ❖ CPUs adopt part of the ICTN ISA
 - (they did it with synchro LL&SC, MOESI , MPEG etc. - why not for comm?)
 - also, CPUs integrate a multi-threaded comm. assist / switch-router;
 - eventually P, M and IO communicate natively & explicitly via ICTN transactions;
- 3. I²N - Intelligent networks => prg-able and adaptive ICTNs