

Paper this morning about cost / performance of cluster interconnects

you need either...

- a **radical commodity** approach using the **cheapest possible components** to build a *Beowulf*...

... or ...

- a **full custom interconnect** together with a really carefully architected system using all tricks (e.g. SMP nodes) to build a large high end machine like the *PSC Terascale*...

... to get a good cost / performance ratio.

Top three limiting technologies?

The IO Bus... (Microchannel)

The IO Bus... (PCI)

The IO Bus... (PCI-X)

All past and current **I/O buses** were and are still **inadequate** to interface a superb **fast interconnect** to a good **high speed CPU**.

- **Communication Co-Processors** do not help much, because I/O busses are equally ill suited to interface a main-processor with a co-processor.
- All the badly needed **memory coherency** signals are still missing. So it is impossible to make a network interface that interacts in an intelligent way with the **local memory systems**.

Features and improvements needed?

- Support for a **global address** space (full coherence optional!).
- Support for **fine grained data communication** (sparse matrices).
- Support for **fast synchronization** (barriers).
- Support for **different traffic classes** (logical channels).

All things we found in early parallel computer like a Cray T3D...

... network interfaces and fast switch ports at a cost that we can afford in a cluster without tricks (like SMPs as nodes)

- The **10x at a time** speed increase of the networking world is ill suited (we need to get 2x, 4x and 8x improvements quicker).
- The trend towards **more features**, e.g. more security, virtuality & administrative features networking equipment is **detrimental**.

We just need basic functions and high performance.

Complete design freedom would lead to...

A new high performance microprocessor with integrated network interface with:

- Multi threading for different communication activities.
(e.g. sending, receiving, short messages, bulk messages)
- Efficient primitives for synchronization (all and partial cluster).
- Adequate mechanism for bulk transfers.

Communication will ultimately become (again) a first order construct in programming languages.

- Direct statements and instructions for moving data in a distributed machine, no more borrowing of memory semantics!

A hardware software co-design including a new OS.

- Complexity - no modifying the LINUX or WinNT kernels.

Why are interconnects still an excellent research opportunity?

- **Golden Age** of MPP Architecture (1985-1995)
Microprocessor with a network interface integrated into the register file! iWarp (CMU, DARPA-Intel) DASH, T3D, T3E.
- **Dark Age** of Interconnect Architecture (1995-2005)
Mandatory use of a broken IO Bus to keep up with Microprocessor generations, Microchannel or PCI.
- **New Age** of Interconnect Architecture (2005-)
CPU Vendors will finally add a **Network Access Port** to their Microprocessor chip-sets - just like AGP to make fast graphics happen.

With an network port the communication systems architects will no longer be second class citizens!