Is Exascale the End of the Line for Commodity Networks?

Scott Pakin

Applied Computer Science Group

Los Alamos National Laboratory

27 September 2011

Short Answer: No

***Disclaimer*: The opinions expressed herein do not necessarily reflect the positions of Los Alamos National Laboratory, the National Nuclear Security Administration, or the United States Department of Energy …or, for the most part, me.**

Argument #1: Commodity Networks Worked at Petascale

- **Roadrunner @ Los Alamos**
  - First sustained petaflop/s
  - 3,060 nodes of InifiniBand
  - First Top 1 supercomputer ever to use a commodity network
- **No multi-PB/s optical data vortex with cryogenic light sources**
  - Sounded like a good idea for petascale back in 1999

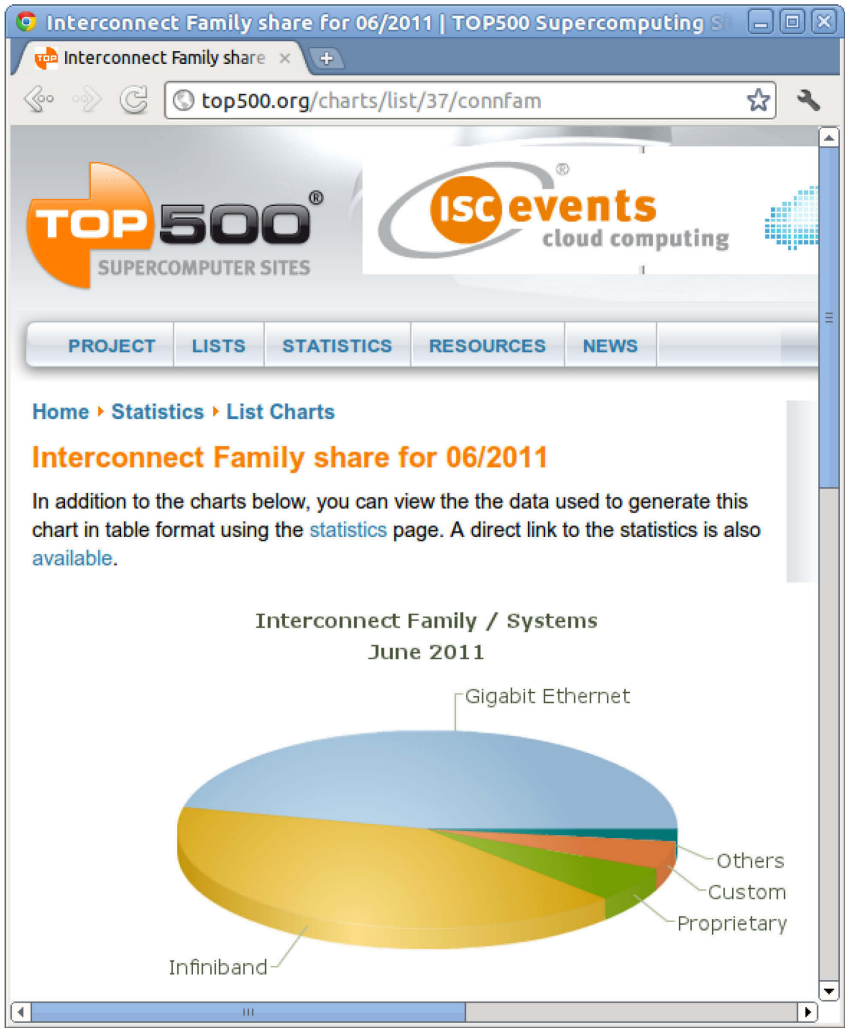
- **Exascale possibility #1**
  - Custom networks within a compute unit (e.g., a rack)
  - Commodity network interconnects the compute units
  - Not all that different from ASCI Blue Mountain, ca. 1999 (SGI NUMAlink intra-node, commodity HiPPI inter-node)
  - For concreteness, consider, e.g., a Blue Gene-like system of 326 IB-connected racks, 1,024 sockets per rack, and 3 Tflop/s GPUs instead of low-end CPUs

Argument #2: Cost

- **Underlies almost every argument for using almost any commodity**
- **Get something "good" for significantly less money than "perfect"**
  - I could buy a custom-tailored suit that fits perfectly and looks exactly the way I want
  - The clothes I'm wearing now fit fine, look okay, and cost significantly less
- **Leaves more money to spend on other parts of the system**

Argument #3: Misguided Performance Metrics

- **Does "exascale" mean "10^{18} flop/s on LINPACK"?**
  - Metric for sorting the Top500 list
  - People who pay for really big supercomputers like to see them in the #1 slot
- **LINPACK transmits only O(N^2) data for O(N^3) computation**
- **Moral**
  - Buy a relatively cheap network
  - Put the money saved into more and faster processors
- **(Oh, you actually wanted to run *applications* at exascale?)**



Argument #4: Fewer Unknown Unknowns

- **Commodity networks have all sorts of problems from an HPC standpoint**
  - Per-connection resource requirements
  - Pre-pinning of communication buffers
  - Bulky routing tables to handle arbitrary topologies
  - Many cycles needed to trigger communication
- **Point is that we know what the problems *are***
  - Academia figures out how to work around most network shortcomings
  - Industry eventually produces great implementations of awful standards
  - Why is my one-off network sometimes slow? Who knows? (Limited experience and few tools)

"[T]here are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know."

Argument #5: Stupid, Meddling Bureaucrats

- **U.S. regulations prohibit granting supercomputer access to a non-U.S. person without acquiring an export license**
- **What's a supercomputer?**
- **If the system uses a *proprietary* network, then**

$$\sum_{i=1}^{n} \underbrace{W_i \left(\frac{FPO_i}{t_i}\right)}_{\text{Node performance in "weighted Tflop/s" (an idiotic metric)}} > 1.5 \text{ WT}$$

- **If the system uses a *commodity* network, then**

$$\max_{1 \le i \le n} W_i \left(\frac{FPO_i}{t_i}\right) > 1.5 \text{ WT}$$

- **References**
  - U.S. Dept. of Commerce. *A Practitioner's Guide to Adjusted Peak Performance*. Dec. 2006
  - U.S. Export Administration Regulations, Part 774: Commerce Control List, Category 4 (Computers), Supplement No. 1

Conclusion

- Let's go build some exascale supercomputers with commodity networks!
- It won't be a horrible mess…really!