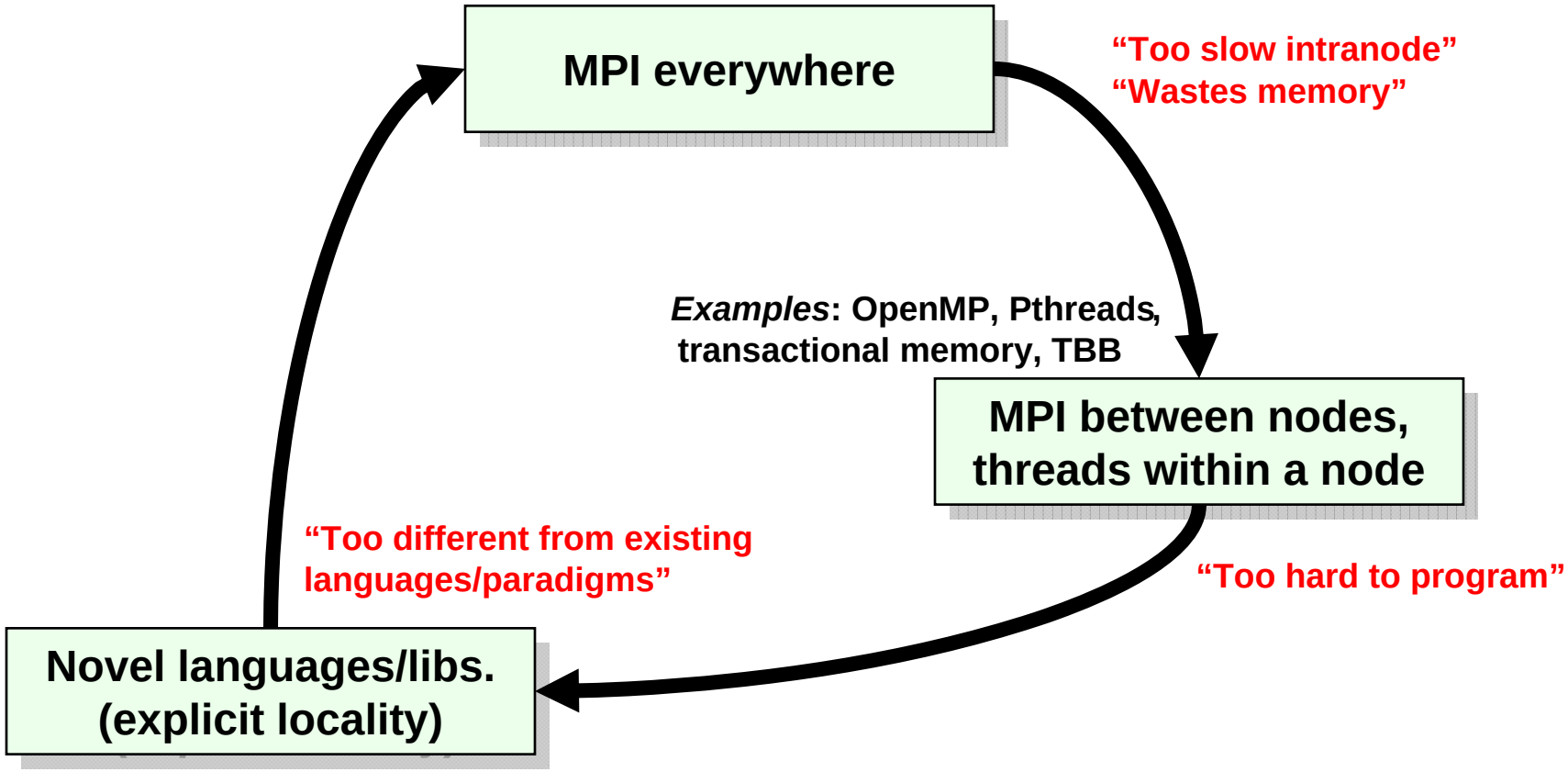
# All I Want From an Interconnect Is…

**Scott Pakin**

**Performance and Architecture Lab (PAL)**
**Los Alamos National Laboratory**

**21 July 2008**

# Programming Models: Past, Present, and Future




*Examples*: UPC, CAF, ARMCI, Fortress, Chapel, X10, Titanium

# New Challenges = Old Challenges

- **Think of network as next level of memory hierarchy**
  - Want no worse than 10X drop in performance from previous level
  - Same metrics are important
- **Regardless of programming paradigm:**
  - Small amount of global data can be kept locally
  - Must communicate data between levels of memory hierarchy

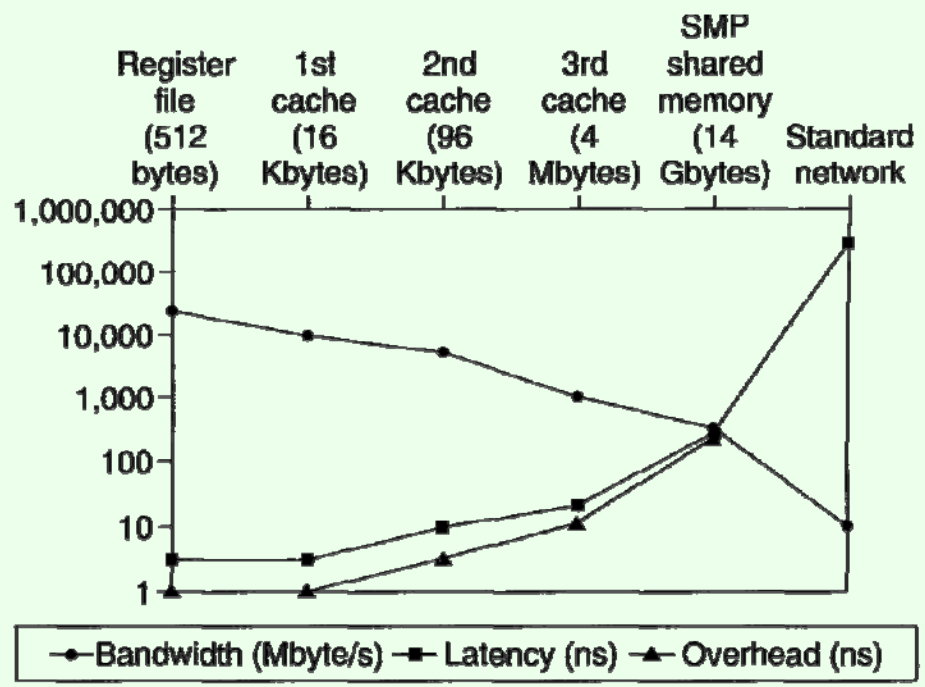


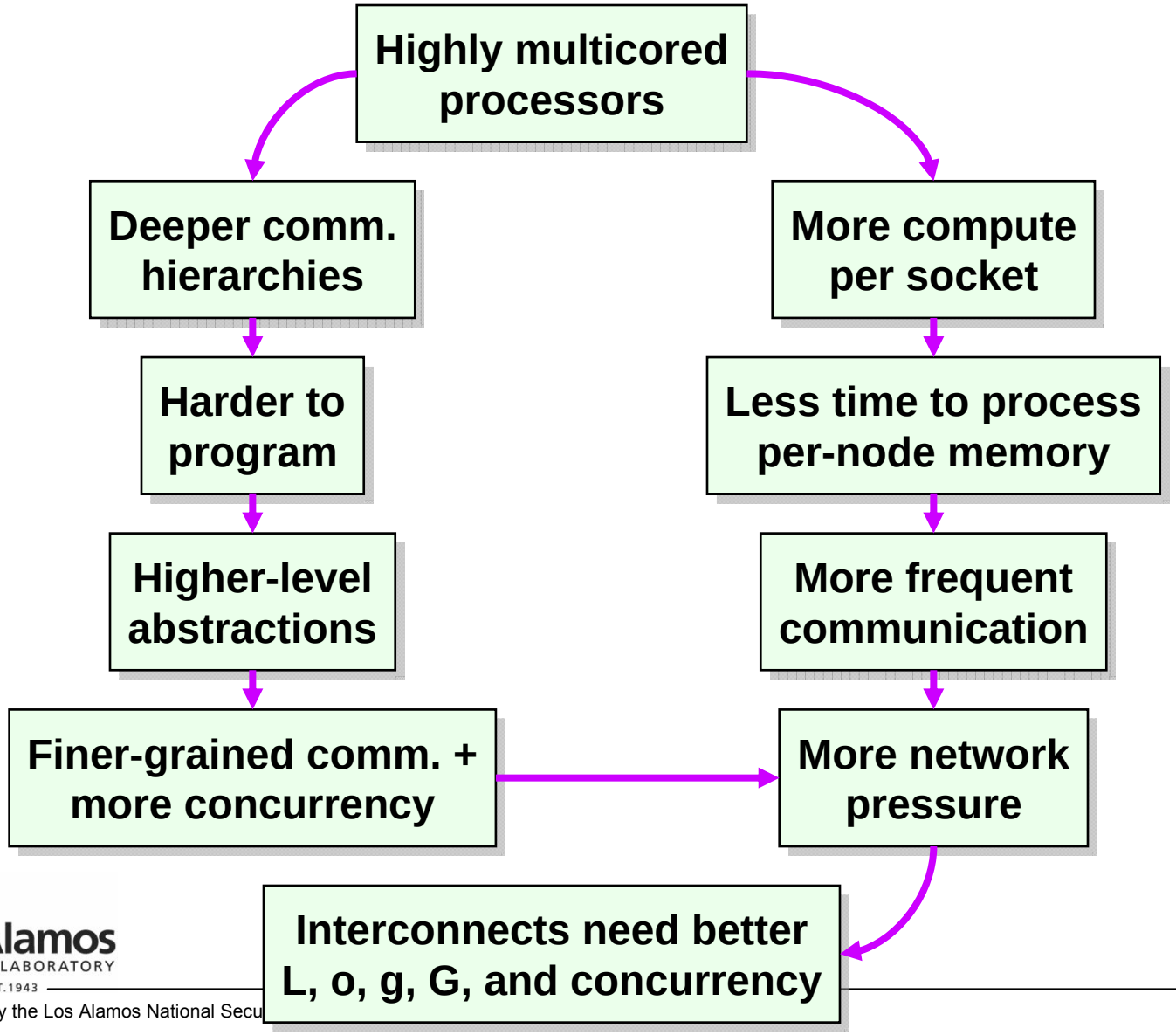

Figure 1. Digital AlphaServer 8400 memory hierarchy communication characteristics (SMP: symmetric multi-processor system).

R. Gillett, "Memory Channel Network for PCI". *IEEE Micro*, Feb. 1996.

# New Challenges ≠ Old Challenges

- **Deeper comm. hierarchies, and getting deeper still**
- **Roadrunner**
  - SPU registers → local store → Element Interconnect Bus → Broadband Interface → Cell-blade memory (XIO) → PCI Express → per-Opteron memory → HyperTransport → InfiniBand
  - Cell and Opteron memories and InfiniBand network are themselves hierarchies
  - {SPU, PPU, Opteron} × {SPU, PPU, Opteron} = many hierarchies
- **Qualitative difference from the past**
  - Impure bandwidth hierarchy
  - Soda straws connecting fire hoses
- **Qualitative similarity**
  - Locality = good, randomness = bad

# Implications for Interconnects



# Characteristics of a Good Interconnect

- **Overprovisioned performance**
  - Key is ratio of interconnect speed to compute speed
  - As network speed→∞, more parallelism becomes beneficial
  - (Would you parallelize $(a+b)\cdot(c+d)$ across two processors?)
- **Low communication overhead**
  - Enables hiding latency behind concurrency
- **Simple cost model**
  - Time proportional to message size (want $t(s)=t_0+s/r_\infty$)
  - Latency and bandwidth degrade predictably with distance
- **Facilitate programming-model implementations**

# Summary

- **Specific programming model is largely irrelevant**
  - Data that doesn't fit locally requires network communication
- **Interconnect needs to be fast *relative* to compute**
  - RDMA, remote atomics, scatter/gather, etc. useful when they enable more concurrency to translate to more performance
  - Not so useful if they complicate performance characteristics
- **Make it easy to reason about network performance; let programming models do the rest**