Scaling Laws, Multicore, and GEANT4 Optimization Opportunities.

Rob Fowler
Renaissance Computing Institute (RENCI)
University of North Carolina

May 8, 2012

Context:

- At Lali Chatterjee's request, researchers from SciDAC-2 PERI took a look at GEANT4 for optimization opportunities.
  - PERI's initial charter was on shorter term opportunities.
    1. log() and exp() dominate in calorimeter examples.
    2. Identified trigonometry "low hanging fruit" in magnetic fields.
    3. In all examples, the C++ method call stack is very deep.
    4. → good IPC rates, #3 raises overhead vs "real work" issues.
  - Broader issue: tension between human productivity vs. machine efficiency.
    - What objective function to use? Who decides?
    - Can we have both? *E.g.*, domain-specific languages, …
- This workshop is broadening the charter
  - Current and near future multi- and many-core machines.
  - Bigger issue: Physical constraints will radically change future systems.
    - Example: Dramatic changes needed for exascale systems.

The Memory Wall

- "Hitting the Memory Wall – Implications of the Obvious." Wulf and McKee 1994. ("Reflections on the Memory Wall", McKee, CF'04)
  - —Processor speeds (clocks) were increasing exponentially.
  - —Memory speeds increasing with much smaller exponent.
  - → Improvements to caches will only stop the bleeding temporarily.
  - → Fundamental changes in computer architectures are needed.
- *"It's the Memory, Stupid", Richard Sites, DEC (in Microprocessor Report, 1996)*
- "(for a certain set of applications) processor speed is already effectively infinite compared to memory. The only relevant benchmark is Stream." – Bob Morgan, DEC (~2001).
- What's changed, if anything? Processor clock speed has leveled, but now we have multi-core and multi-threading.

Dennard Scaling of CMOS Logic.

- Series of papers 1972-1974 by Bob Dennard and others at IBM on scaling properties of CMOS logic circuits (gates and wires!).
- Linear scaling of all transistor parameters.
  - Reduce feature size by a factor of S,typically .7/generation.
    - Including gate insulator thickness!
  - Reduce supply voltage (Vdd) by S to keep electric field constant.
  - Adjust doping of silicon gate region to compensate.
  - Area shrinks by S^2, C_{gate} and delay (1/f) reduced by S.
  - Power ≈ CV^2f ➔ Power per gate goes down by S^2
  - Area and power track each other so power density is unchanged.
  - For a constant die area and design density, power and power density are constant and frequency increases.

Other Aspects of Dennard Scaling.

- Wire resistance/unit length ~ S^2
- Wire capacitance/unit length ~ 1
- RC delay/unit length (unrepeated) ~ S^2
- Die size (D) increases, so "long" wires increase by D
- Unrepeated wire delay ~ S^2D^2, repeated ~ D sqrt(S)

  →Signals cannot cross the chip in one cycle.

Moore's law

Empirical observation and self-fulfilling prophesy:
Circuit element count doubles every N months. (N ~18)

- Technological explanation: Features shrink, semiconductor dies grow.

- Dennard scaling: Gate delays decrease. Wires are relatively longer/slower.
  - Dennard scaling has not been perfect in practice and is coming to an end.

- In the past, the focus has been making "conventional" processors faster.
  - Faster clocks
  - Clever architecture and implementation → instruction-level parallelism.
  - Clever architecture (speculation, predication, etc), HW/SW Prefetching, and massive caches ease the "memory wall" problem.
- Problems:
  - Faster clocks --> more power.
  - Power scaling law for CMOS: $P = \alpha C V^2 F$, but $F_{max} \sim V$ so $P \sim F^3$
    - Where α is proportional to the avg. number of gates active per clock cycle.
  - Smaller transistors + long wires → either slow clock, or pipelined communication.
  - More power goes to overhead: cache, predictors, "Tomasulo", clock, …
  - Big dies --> fewer dies/wafer, lower yields, higher costs
  - Aggregate effect --> Expensive, power-hog processors on which some signals take 6 cycles to cross.

The End of Dennard Scaling and Dark Silicon

- Vdd Scaling issues
  - Initially, designers constrained by standards: 12V, 5V, 3.3V.
  - On-board power regulation now allows Vdd to be 1V or less.
  - This is getting uncomfortably close to threshold voltages.
  - Decreasing thresholds has rapidly increased leakage current/power.
  - Decreasing f allows operation with higher thresholds.
- Gate Insulator issues
  - Thickness is now ~ 5 atoms
- Useful work and duty cycles
  - Bailey and Snyder (1988) observed that a was at most a few percent for processors. If a were much larger, chips would melt.
  - Aggressive architectures have increased a to do bookkeeping, data movement, …
- "Dark" and "dim" silicon refer to schemes to reduce a and/or f to reduce power.
  - "Turbo" modes actually throttle f when all cores are active.
    - Run power-efficient, low f, low V in highly parallel code regions.
    - Inefficient high f, high V on few cores in sequential regions.
  - Heterogeneous cores and purpose built modules w. power mangement.
  - Programmable logic and reconfigurable devices.

Little's Law and Memory.

- Classic law/lemma in queuing theory
  - (mean # in system/queue) = (arrival rate) (mean residence time)
- Communication (memory) restatement
  - (concurrency) = (bandwidth) (latency)

→ To increase bandwidth without decreasing latency, you have to increase the concurrency of the system

  - Wider channels to send more bits per operation.
  - Overlapping, i.e., pipelined, operations.

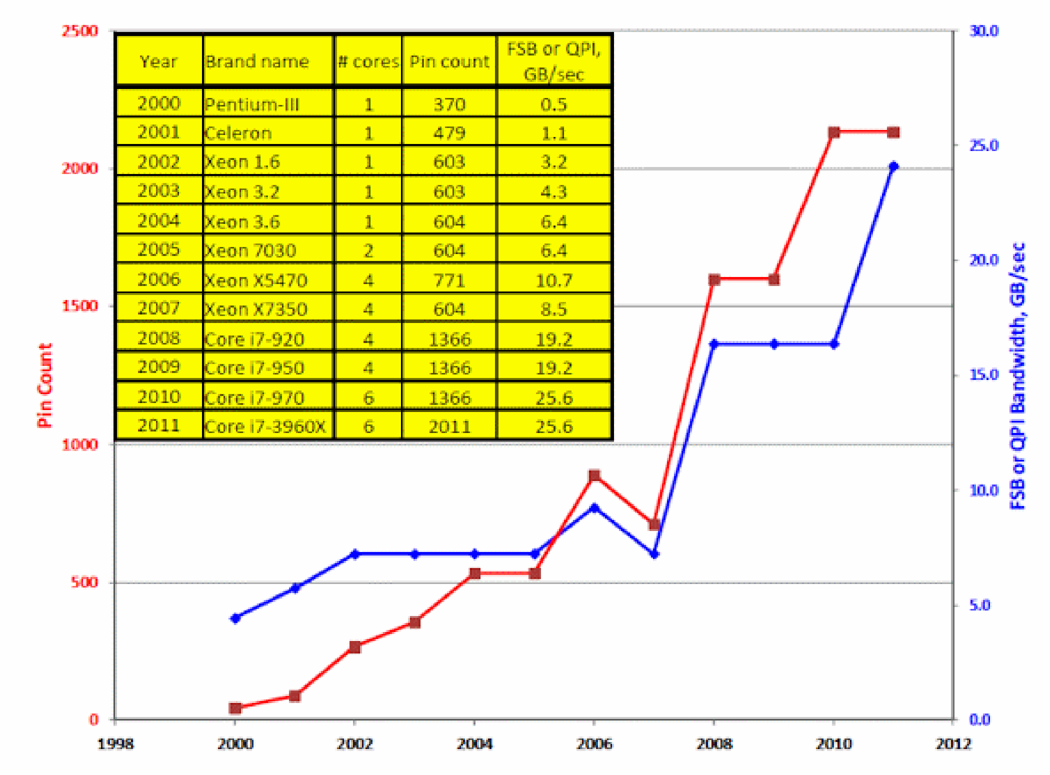
Bottleneck → bandwidth plateaus, queuing latency dominates.

Moore's Law/Dennard Scaling Revisited for DRAM.

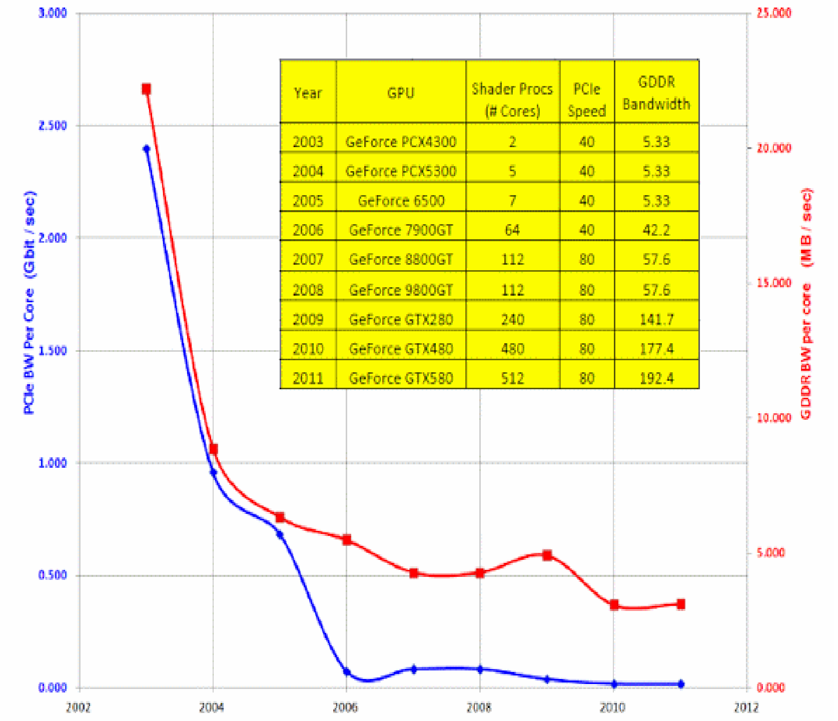
- As more transistors were added to processor chips, they got a lot faster.
  - Clever architectures and on-chip concurrency.
  - Technology: Smaller transistors are faster.
- As more transistors were added to memory chips, they got a lot bigger.
  - Cleverness went into reliability, yield, …
  - Small transistors are fast, but weak (can't drive long wires).
  - Little increase in on chip concurrency.
  - Very low Rent's law (surface/volume ratio) exponent!

	Introduction	Size	Pins	Cycle Time	Bandwidth
DDR	2000	2 GB	168	5 ns	3.2 MB/sec
DDR2	2003	4 GB	184	3.75 ns	8.5 MB/sec
DDR3	2007(2009)	16 GB	240	5 ns	12.8 MB/sec
DDR4	2012(?)				25.6(?) MB/sec

Other Trends: Pins and GPU Memory



Year	Brand name	# cores	Pin count	FSB or QPI, GB/sec
2000	Pentium-III	1	370	0.5
2001	Celeron	1	479	1.1
2002	Xeon 1.6	1	603	3.2
2003	Xeon 3.2	1	603	4.3
2004	Xeon 3.6	1	604	6.4
2005	Xeon 7030	2	604	6.4
2006	Xeon X5470	4	771	10.7
2007	Xeon X7350	4	604	8.5
2008	Core i7-920	4	1366	19.2
2009	Core i7-950	4	1366	19.2
2010	Core i7-970	6	1366	25.6
2011	Core i7-3960X	6	2011	25.6



Year	GPU	Shader Procs (# Cores)	PCIe Speed	GDDR Bandwidth
2003	GeForce PCX4300	2	40	5.33
2004	GeForce PCX5300	5	40	5.33
2005	GeForce 6500	7	40	5.33
2006	GeForce 7900GT	64	40	42.2
2007	GeForce 8800GT	112	80	57.6
2008	GeForce 9800GT	112	80	57.6
2009	GeForce GTX280	240	80	141.7
2010	GeForce GTX480	480	80	177.4
2011	GeForce GTX580	512	80	192.4

Implications of variations of Moore's law

- Memory-bound applications will not benefit nearly as much as the CPU-bound in commodity configurations.
- To match core concurrency, lots of memory parts need to be configured in order to get enough pins and memory buffers.
- Lots of big memory parts → huge memory servers.
- System cost is increasingly dominated by memory cost.

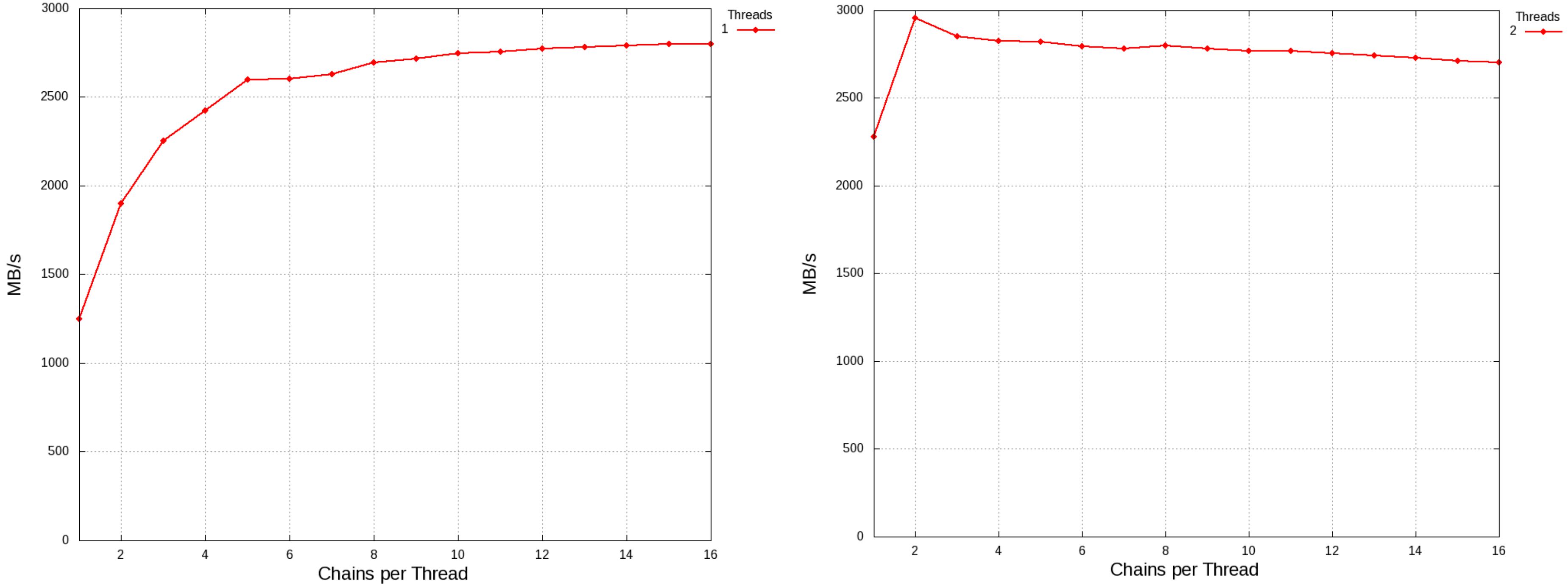
Characterizing Memory Performance

- Most characterization methods use two measures
  - Memory latency (for an isolated operation)
  - Memory bandwidth (for a streaming benchmark kernel)
- ‘STREAM’ and ‘lmbench’ benchmarks – widely used to measure these
- These are often treated as scalar parameters that are fundamental properties of the system
- For multi-socket, multi-core systems, these parameters only tell a part of the story

pChase

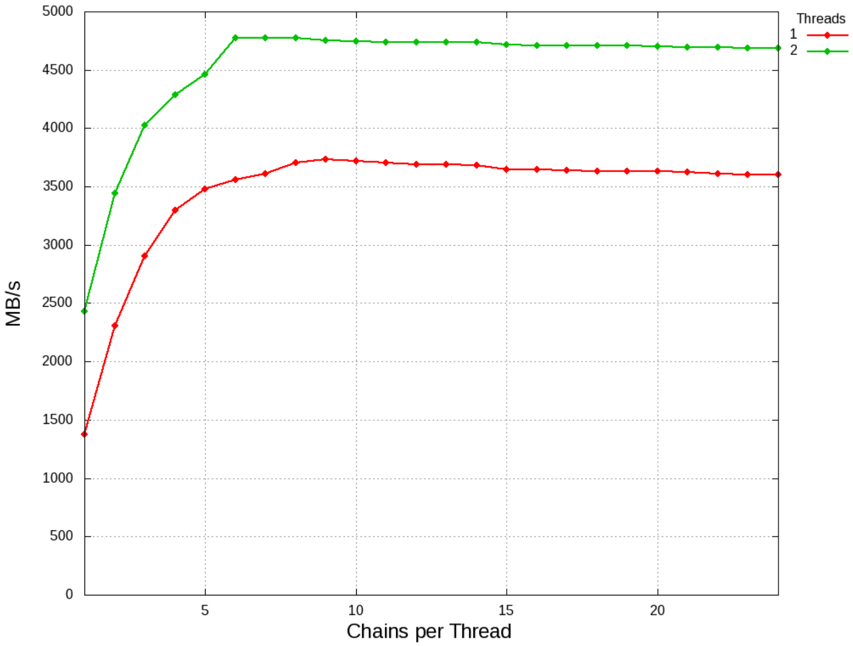
- Developed by Pase and Eckl @IBM
- Multi-threaded benchmark used to test memory throughput under carefully controlled degrees of concurrent accesses
- Each thread executes a controllable number of ‘pointer-chasing’ operations – a memory-reference chain
  - Pointer to the next memory location is stored in the current location. Grow and randomize chain to defeat cache, prefetch.
  - Dereference pointers in k independent chains concurrently, then use them.
- K=1 case measures memory latency.
- Large-k bandwidths are comparable to STREAM measurements at “common” optimization levels.
- Our Modifications
  - Added wrapper scripts around pChase to iterate over different numbers of memory reference chains and threads
  - Added affinity code to control thread and data placement
- Available at http://pchase.org

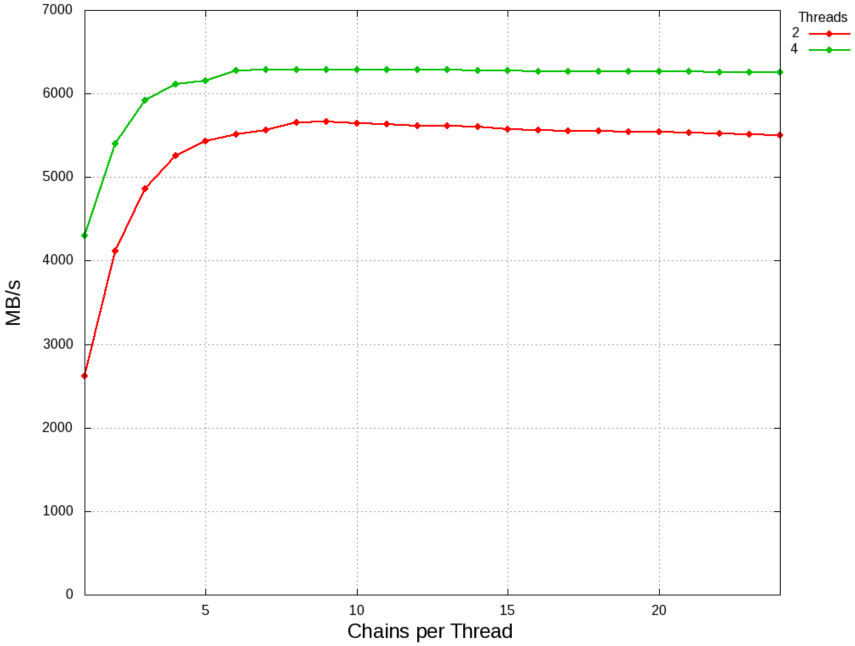
Historical Perspective: ~2004



Dell PowerEdge 1850, 2 x 3.2 GHz Pentium D Xeons
6 x 1 G DDR2 PC3200

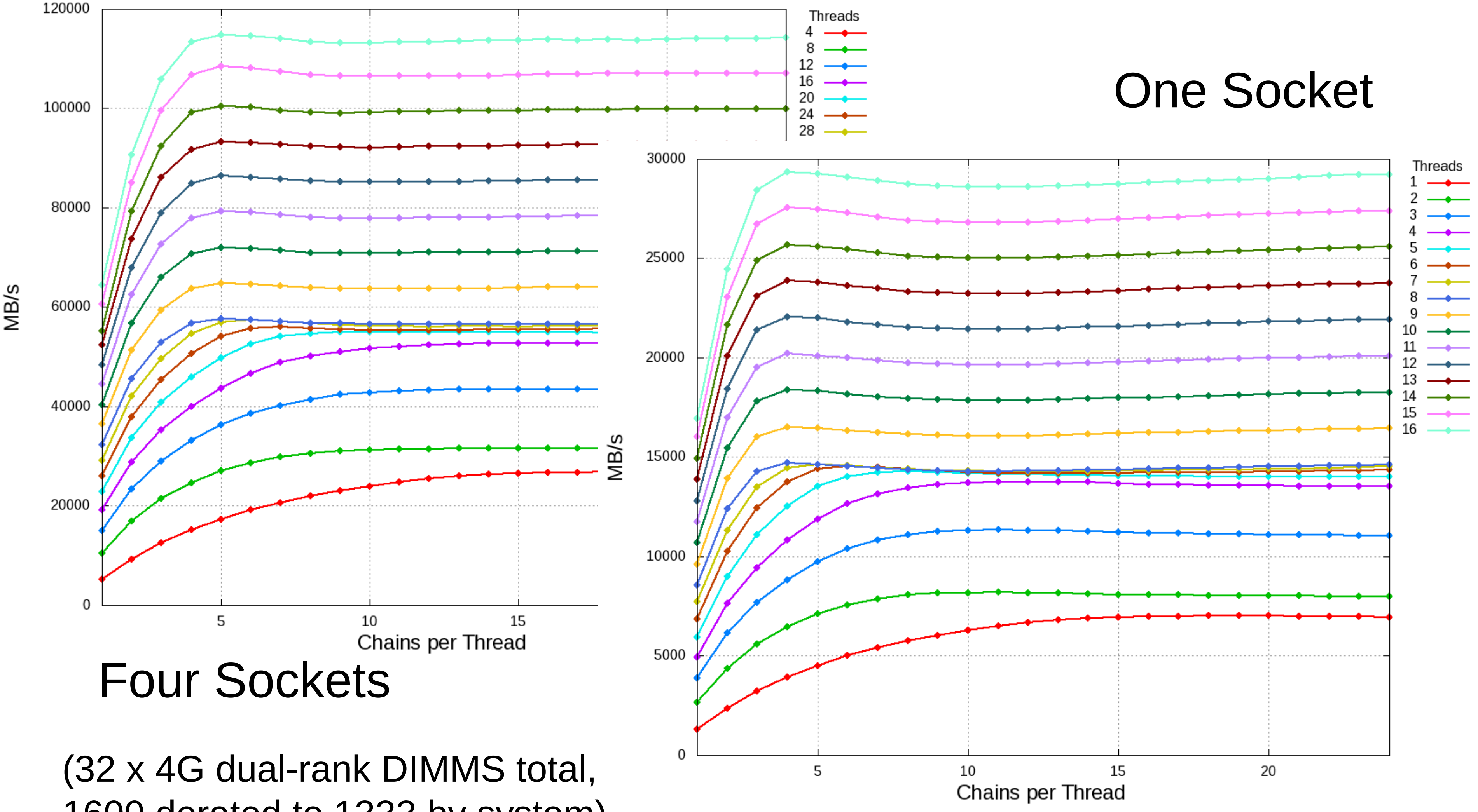
Historical Perspective: ~2006





Dell PowerEdge 1955
2 x Intel X5150, 2 core, 2.66 GHz, (4 cores)
4 x 1GB DDR2 667Mhz

Fully-populated 4-socket Interlagos

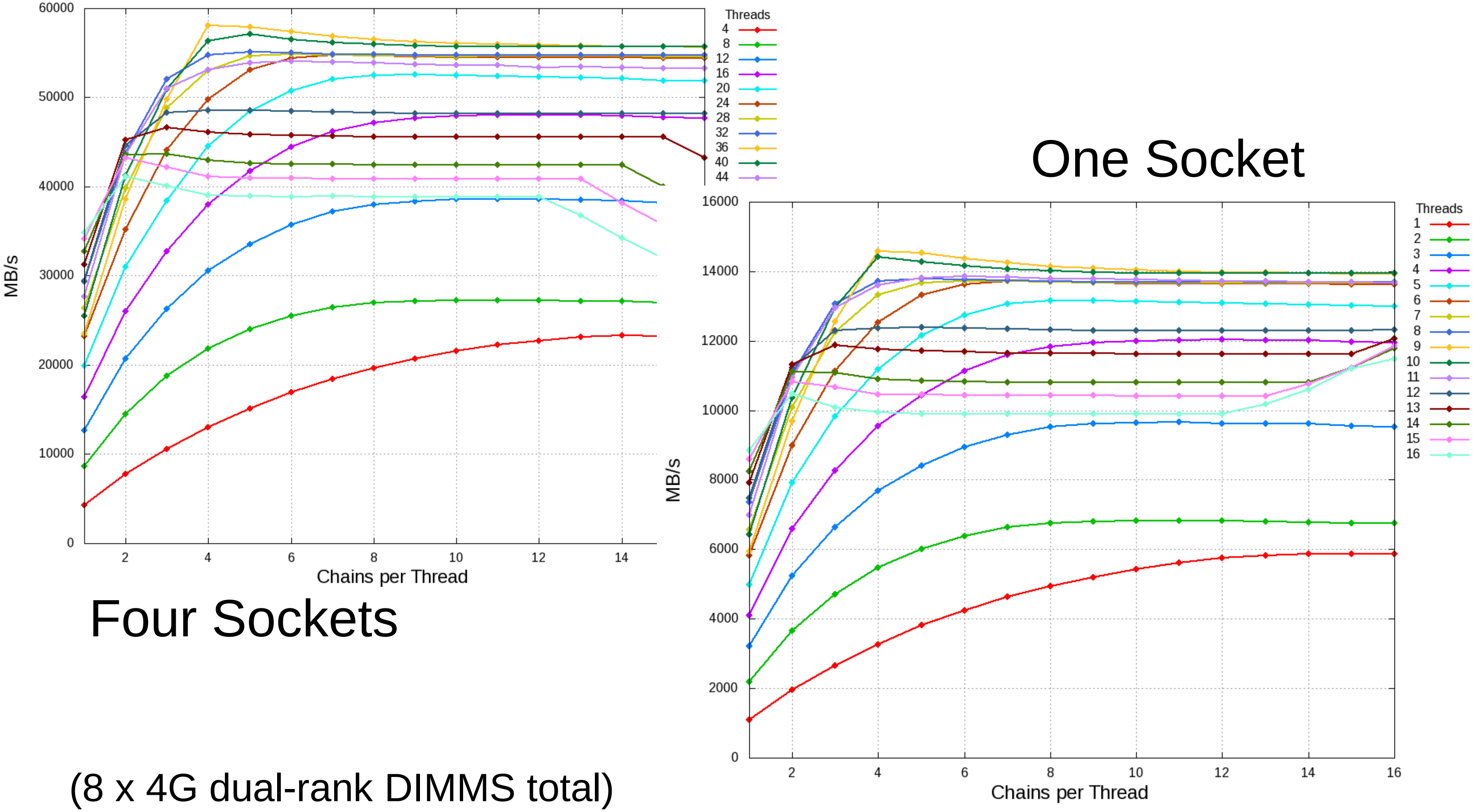


One Socket

Four Sockets

(32 x 4G dual-rank DIMMS total, 1600 derated to 1333 by system)

Interlagos, 2 DIMMs per socket

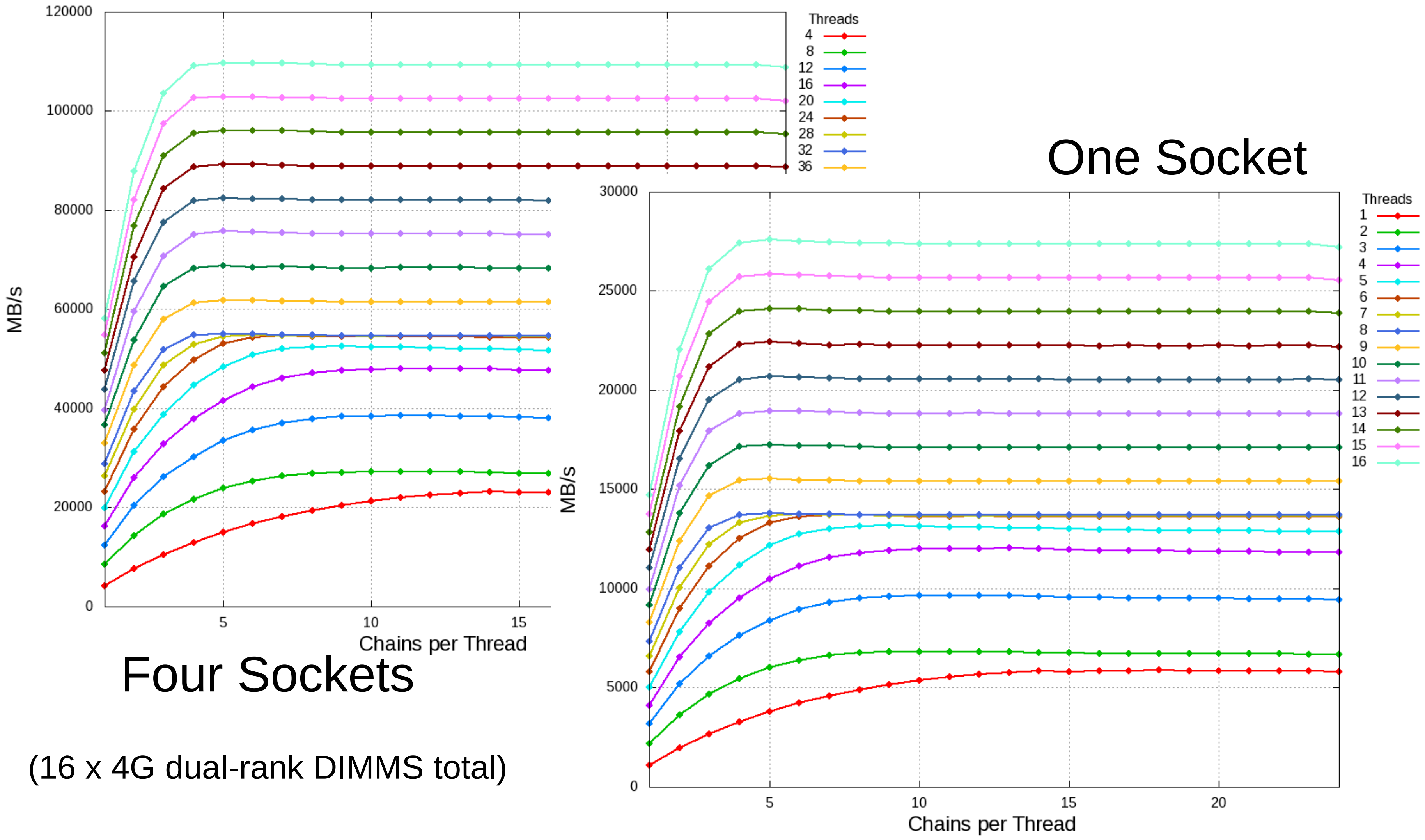


One Socket

Four Sockets

(8 x 4G dual-rank DIMMS total)

Interlagos with 4 DIMMS/socket.

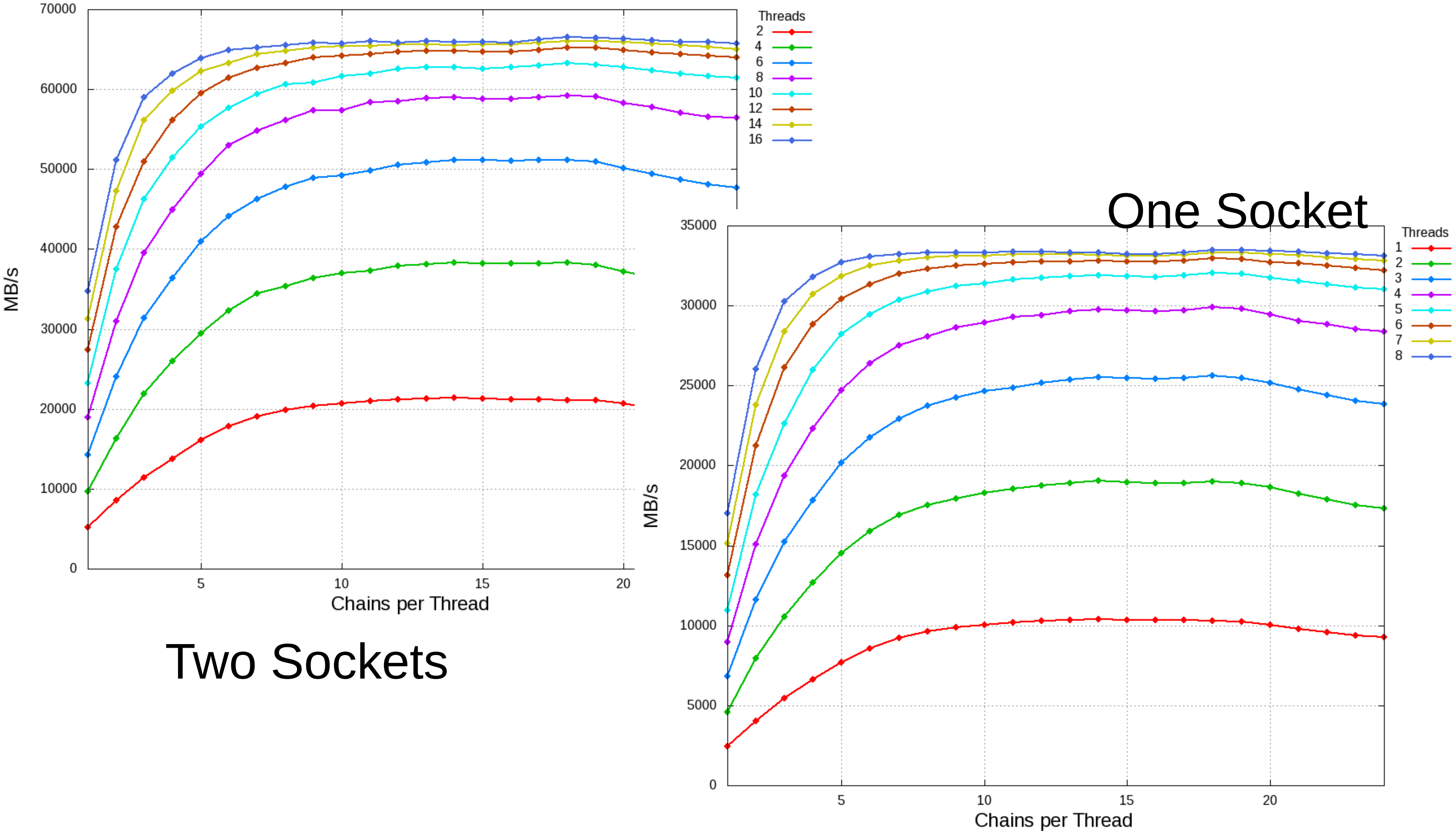


One Socket

Four Sockets

(16 x 4G dual-rank DIMMS total)

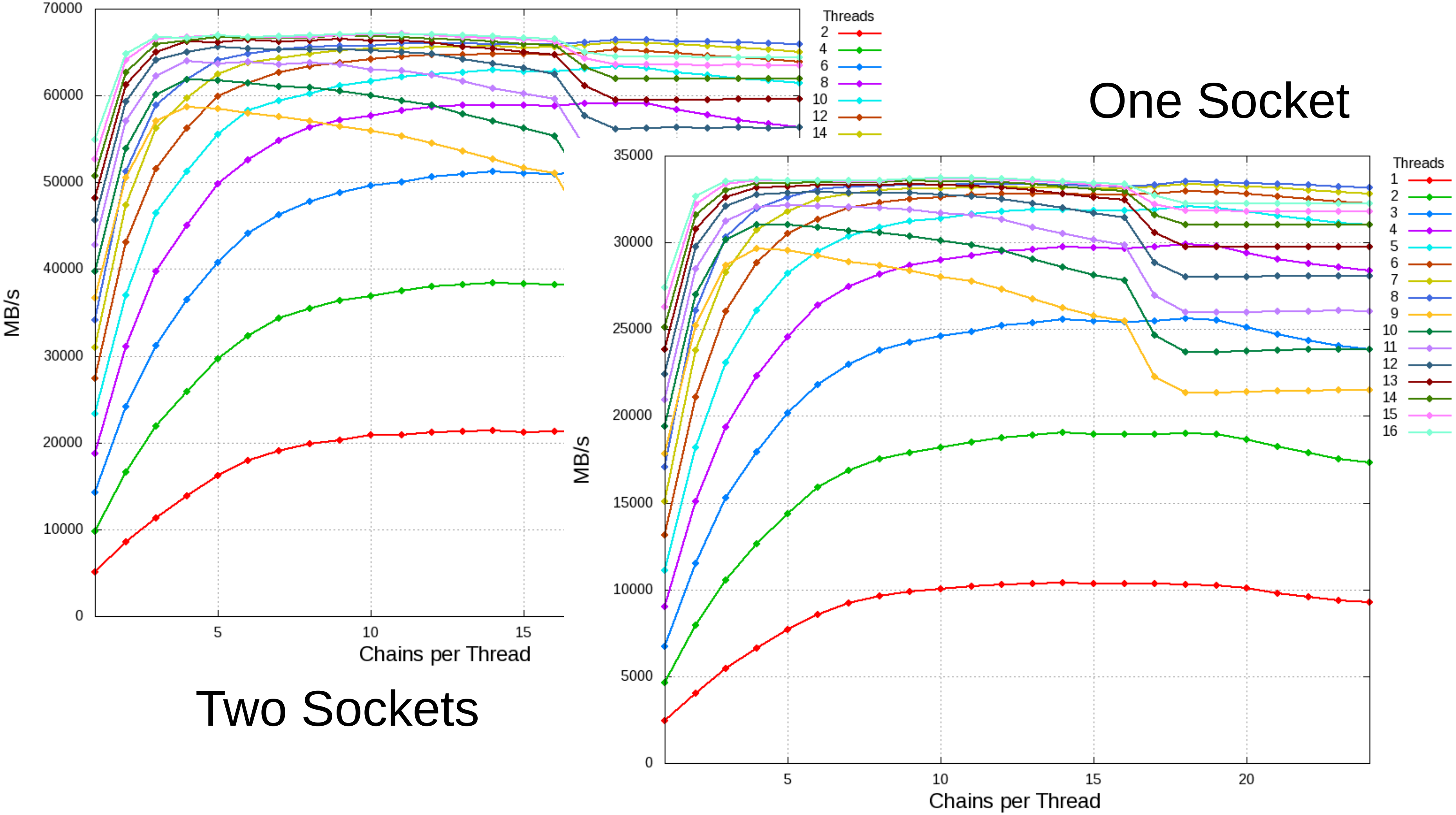
2-socket Sandybridge, w/o hyperthreading



Two Sockets

One Socket

2-socket Sandybridge with hyperthreading



One Socket

Two Sockets

Lessons r.e. Multi-core memory systems

- Per-socket memory bandwidth has increased dramatically.
- "First-core" memory bandwidth has increased even more!
- "Last few cores or threads" incremental memory bandwidth is, in general, poor or non-existent.
- Average per-core (thread) bandwidth has decreased.
  - So has core clock speed if all cores are active!
- Fully-populating all the DIMM slots ($$) on today's high end systems eases the problem.
  - You are buying buffers and interface logic, GBs are a bonus.
  - Do you really need systems with 128 to 512 GB of memory?
    - How much memory do you buy for your 128 core chip?
  - Are you willing to pay for it?
  - What's the business model of processor vendors if memory cost far exceeds the cost of the processor?

Thank you

Contact information:

Robert Fowler (rjf@renci.org)

* **Acknowledgements**
US Department of Defense (DoD)
DOE SciDAC SUPER Institute