Speed and Low Latency Drive Ultra High Speed Messaging and Execution at Wall Street Exchanges

Steve Pawlowski
Intel Senior Fellow
GM, Cross-IAG Architecture and Pathfinding
CTO, Intel Architecture Group

April 4th 2011

intel

The Challenges to Nano-Second Latencies

Network

Data transfer on the network,

Network Hardware Latencies, Network Drivers Latencies, etc

Software

Sync overhead and oversubscription of threads,

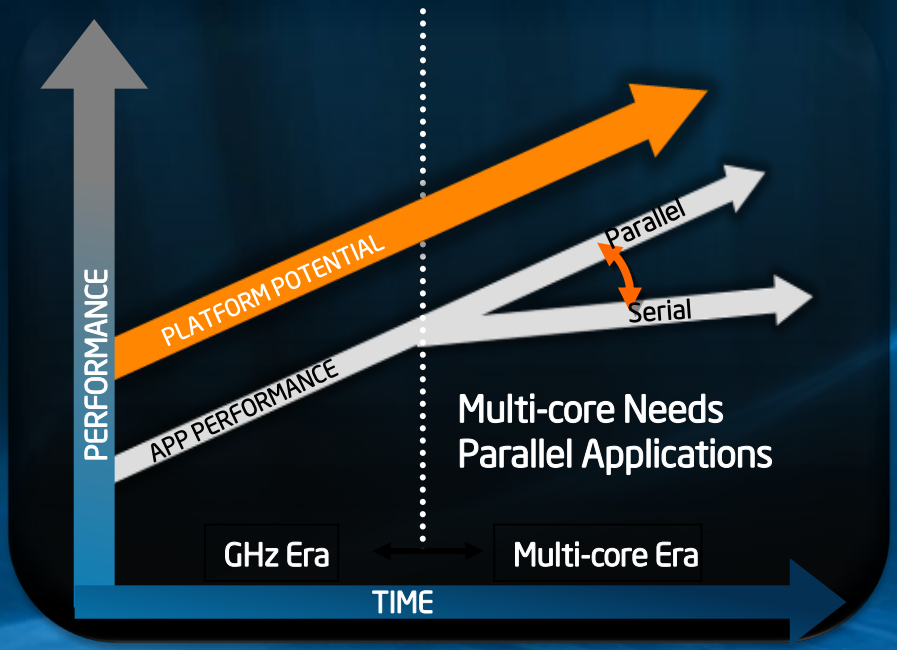
Interrupts, I/O, Parallel Programming, Software Tools, etc

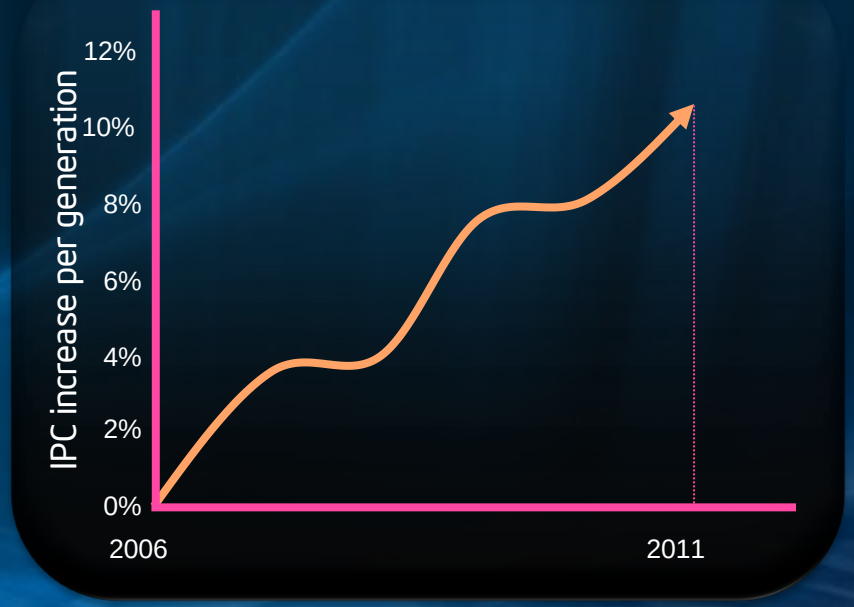
Compute

Frequency, Instructions per cycle,

Cache Sizes, Memory Speed & Bandwidth, RAS

Paradigm Shift:
Era of Multi-threading

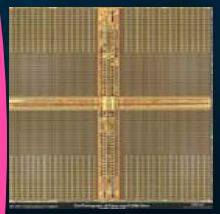




While All PC's Shipped are Multi-Core,
Emphasis Continues on Single Thread Performance

intel

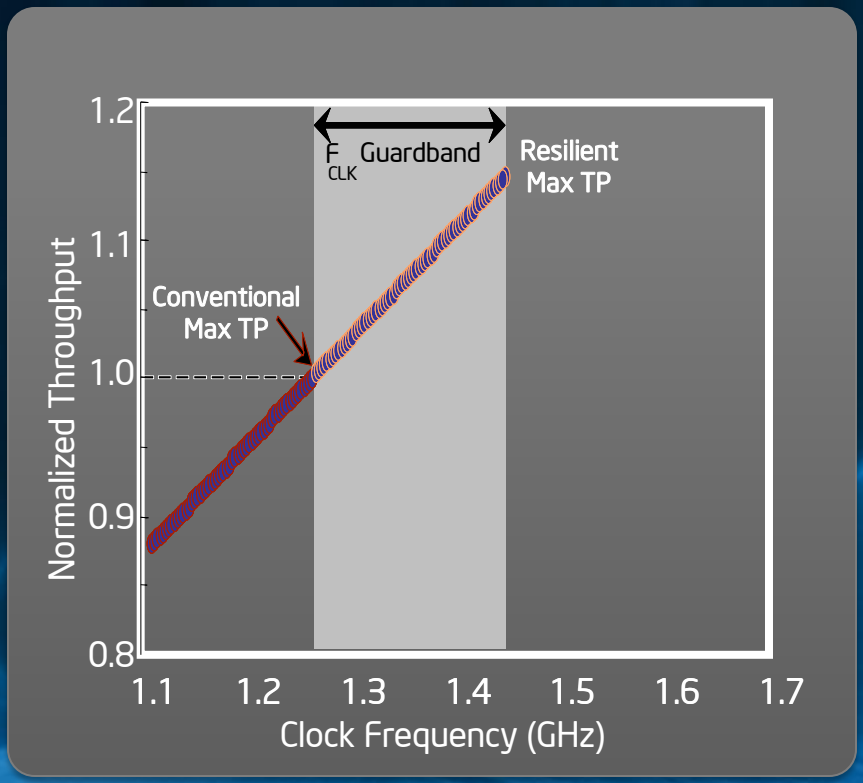
Getting the Data Fast:
A Focus On Memory and Cache Performance

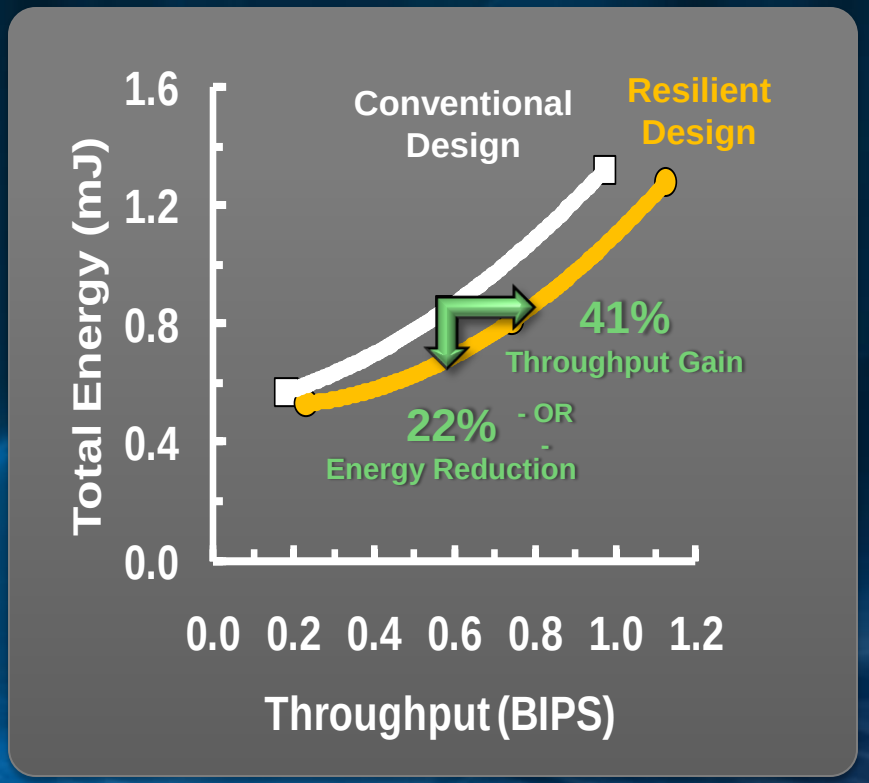


	On CPU SRAM	DRAM	NVM Storage	Storage (HDD)
Latency	1x	~=20x	~=20,000x	~=10,000,000x
$$/bit	1x	~=0.05x	~=0.005x	~=0.0005x

- Continued focus to support higher memory speeds
- Use process technology advances to increase cache sizes and levels of caches
- Investigate new levels of memory hierarchy and wider interfaces
- Architectural modes to reduce cache latency

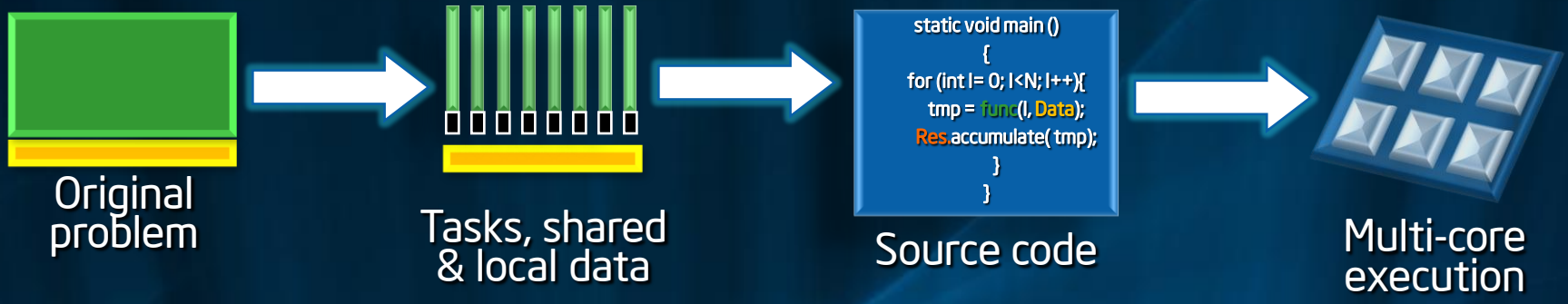
Evaluating the Possibilities with Resilient Designs





- Remove guardbands & use resilient circuits to guarantee correct operation
  - Resilient circuits sense and correct errors from dynamic variations

Parallel Programming Challenges



- *Extracting Concurrency*
  - New Algorithms, Software Tools
- *Expressing Concurrency*
  - Programming Abstractions & Languages + A Change in Mindset
- *Exploiting Concurrency*
  - Compilers, Runtimes & Hardware Support

intel

How Long Until Your Application is Optimized?

Weeks, Months, Years... Alpha to Omega

Focus on Reducing the Time to Optimize...
...Intel CPUs provide support for Legacy Codes...
...Latency of Deployment is As Significant As Any Other Latency

Xeon **Architecture** Philosophy

Both, Highly and Lightly Threaded Codes Matter

Power Efficiency Is Critical

Support for Legacy Codes

Both, Integer and FP Performance Matter

Security

Systems Must Remain Balanced

If we increase the compute density, we must balance that with memory bandwidth

In Conclusion:
The Path to Nano-Second Latencies

Network

Data transfer on the network,
Network Hardware, Network Drivers, etc

- **Minimize distances** between data source and applications
- **Latest networking technologies** such as 10GbE, UDP, RDMA, interrupt moderation, etc

Software

Sync overhead and oversubscription of threads,
Interrupts, I/O, Parallel Programming, Software Tools, etc

- Intel **compiler/performance tools** to characterize/generate software with reduced end-to-end latency
- **Interrupt Affinity** to Cores

Compute

Frequency, Instructions per cycle,
Cache Sizes, Memory Speed & Bandwidth, RAS

- Increased cores, but **focus on single thread** performance
- Increased **cache sizes and hierarchies**
- **Energy efficient** designs
- High end **RAS** features