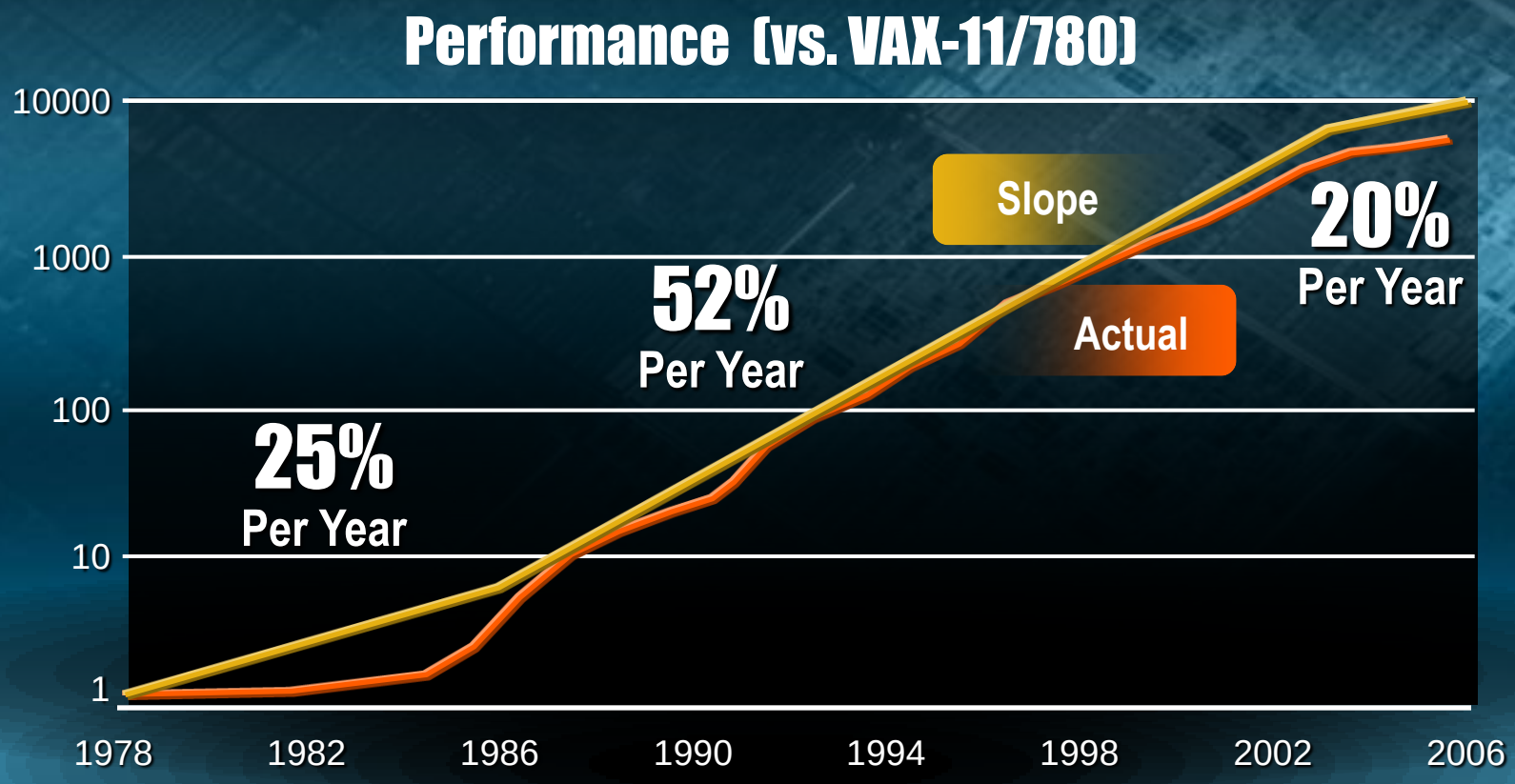

"Micro-architecture
IS DEAD,
LONG LIVE Micro-architecture"
JUSTIN RATTNER
Senior Fellow & Vice President
Chief Technology Officer
Intel Corporation
intel

Slowdown in Single Core Performance Growth



Source - Dave Patterson


What Really Matters?
intel

Locally Instant Response

- People want access to **relevant** digital “stuff” instantly
- Finding the needle in the haystack is what matters
- One device is insufficient
- Content must be accessible across and among devices, seamlessly

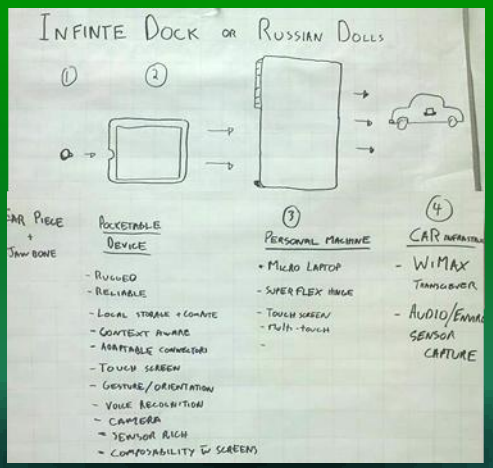
Martin



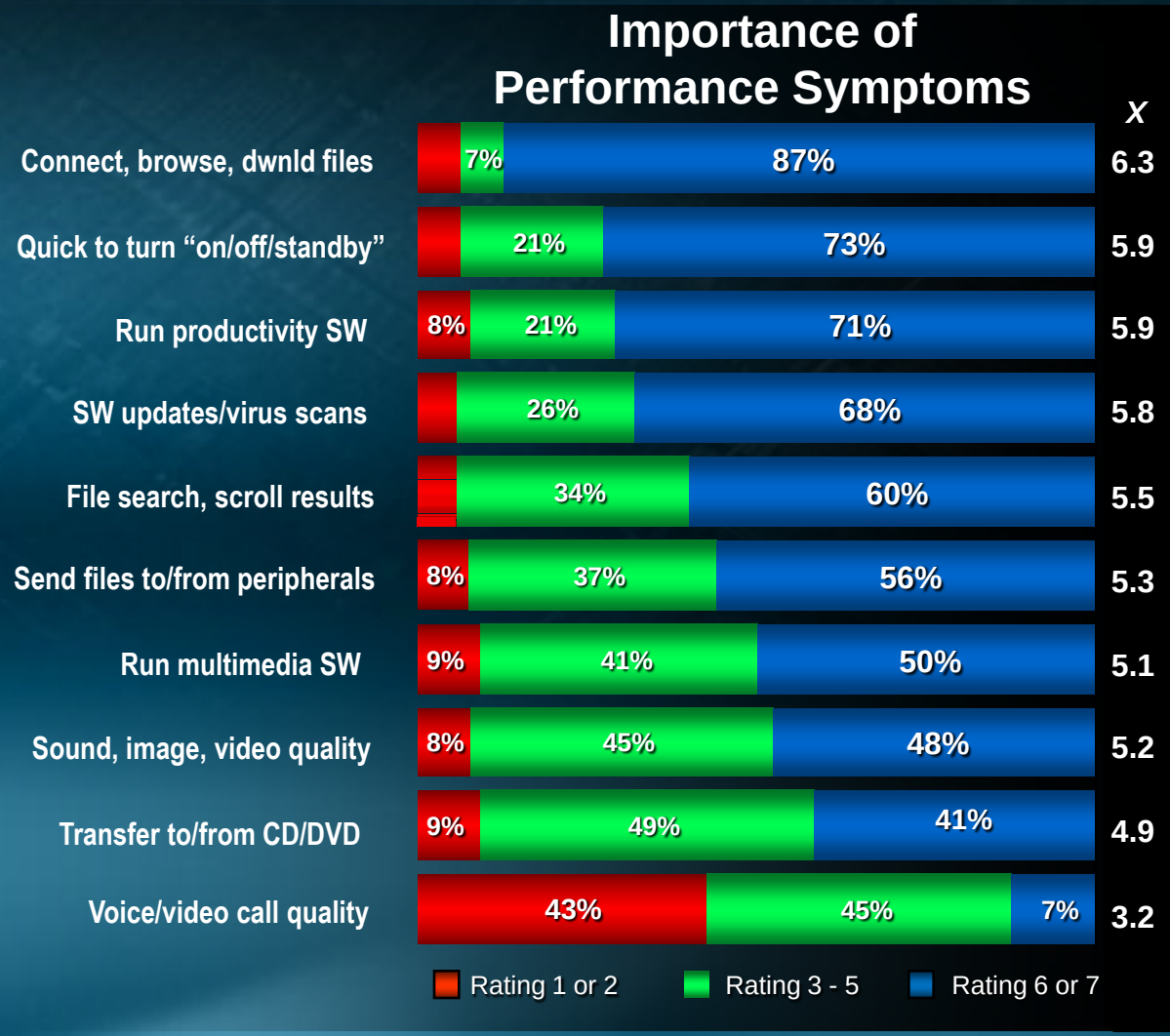
BJ



Darcy



End-User Meaningful Platform Performance



Nearly 90% consider fast Internet usages very important

Nearly ¾ believe quick to 'turn on/off' and 'run productivity SW' is critical

Half of respondents consider multimedia performance very important

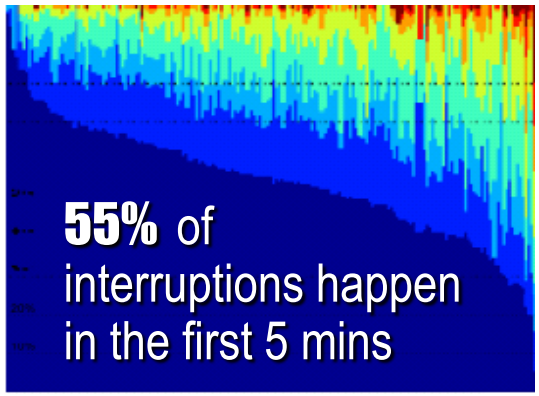
Very few say high quality phone calls on a laptop is essential

Source – Intel PaPR Ken Anderson

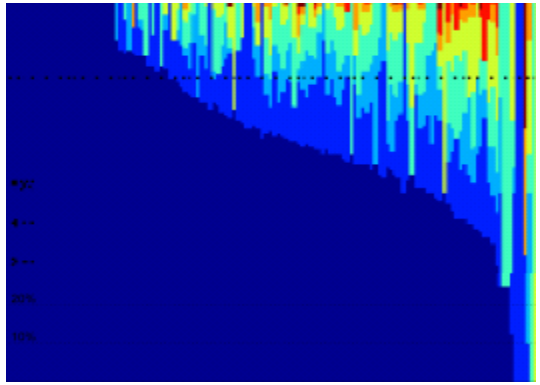
Lots of little hits of IT…

Probability of Interruptions

At Home

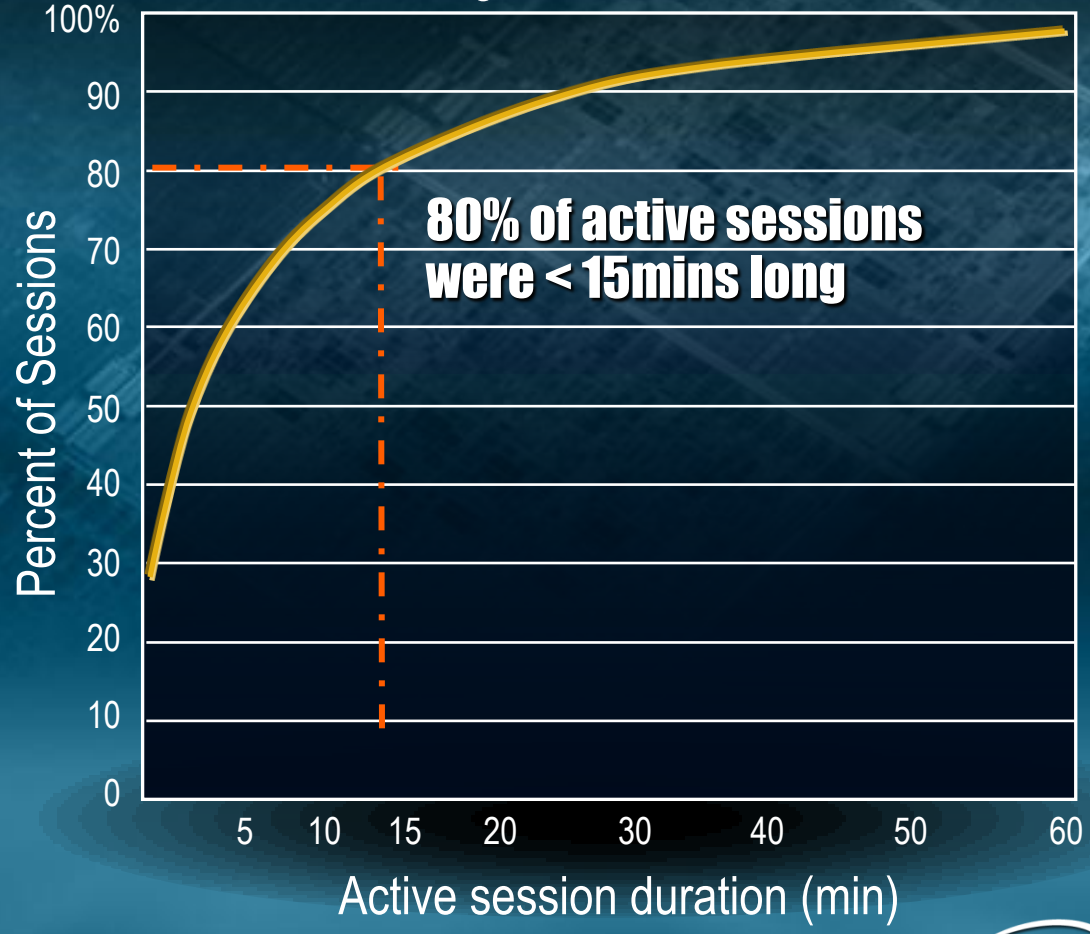


Away



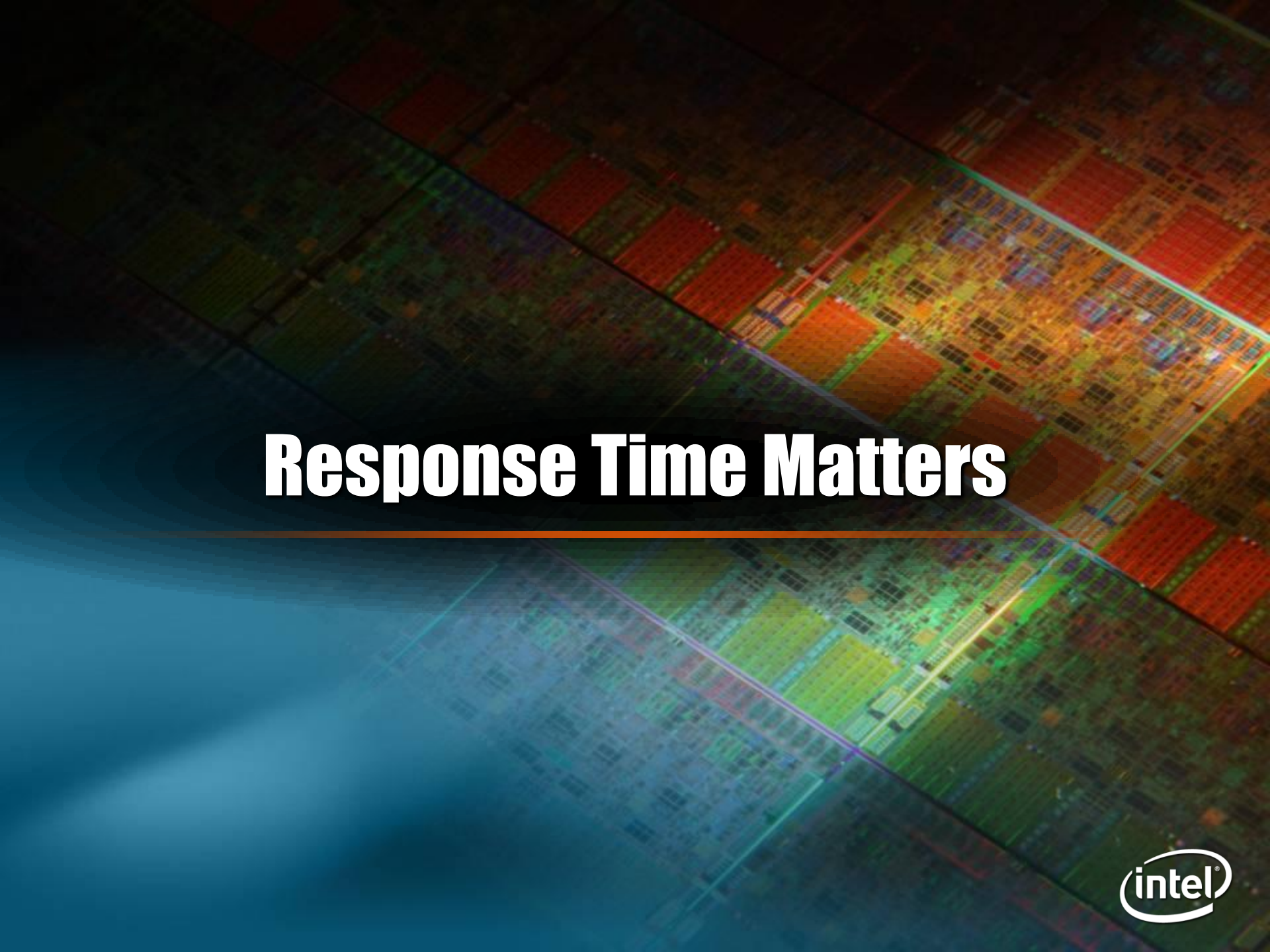
% sessions <= 5 mins
% sessions > 5 mins and <= 10 mins
% sessions > 10 mins and <= 15 mins
% sessions > 15 mins and <= 30 mins
% sessions > 30 mins and <= 60 mins
% sessions > 60 mins and <= 90 mins
% sessions > 90 mins and <= 120 mins
% sessions > 120 mins

Cumulative distribution of average sessions duration



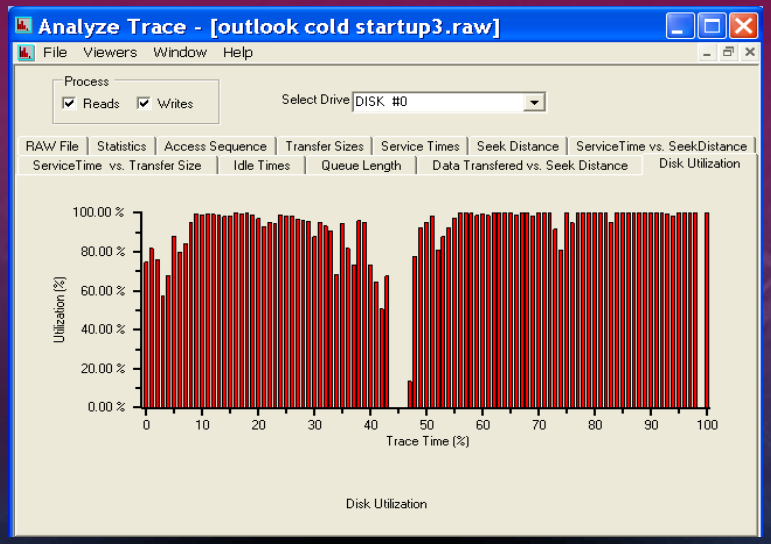
Source - PaPR Ken Anderson, Tye Rattenbury, Dawn Nafus

intel


Response Time Matters
intel

It's the Hard Disk, Stupid!

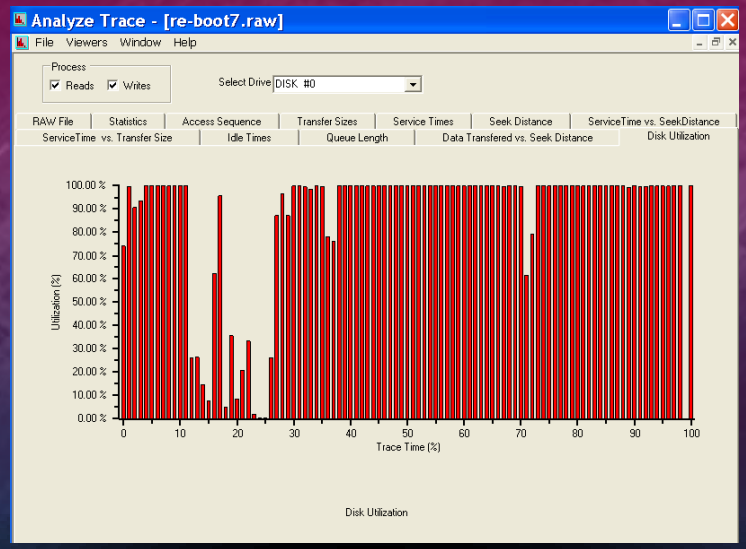
Re-Boot/Startup on Home PC



Elapsed Time	105.213536, s
Disk Busy Time	91.368480, s
Average Data Rate	6.60669, MB/s

86% BUSY

Starting Outlook

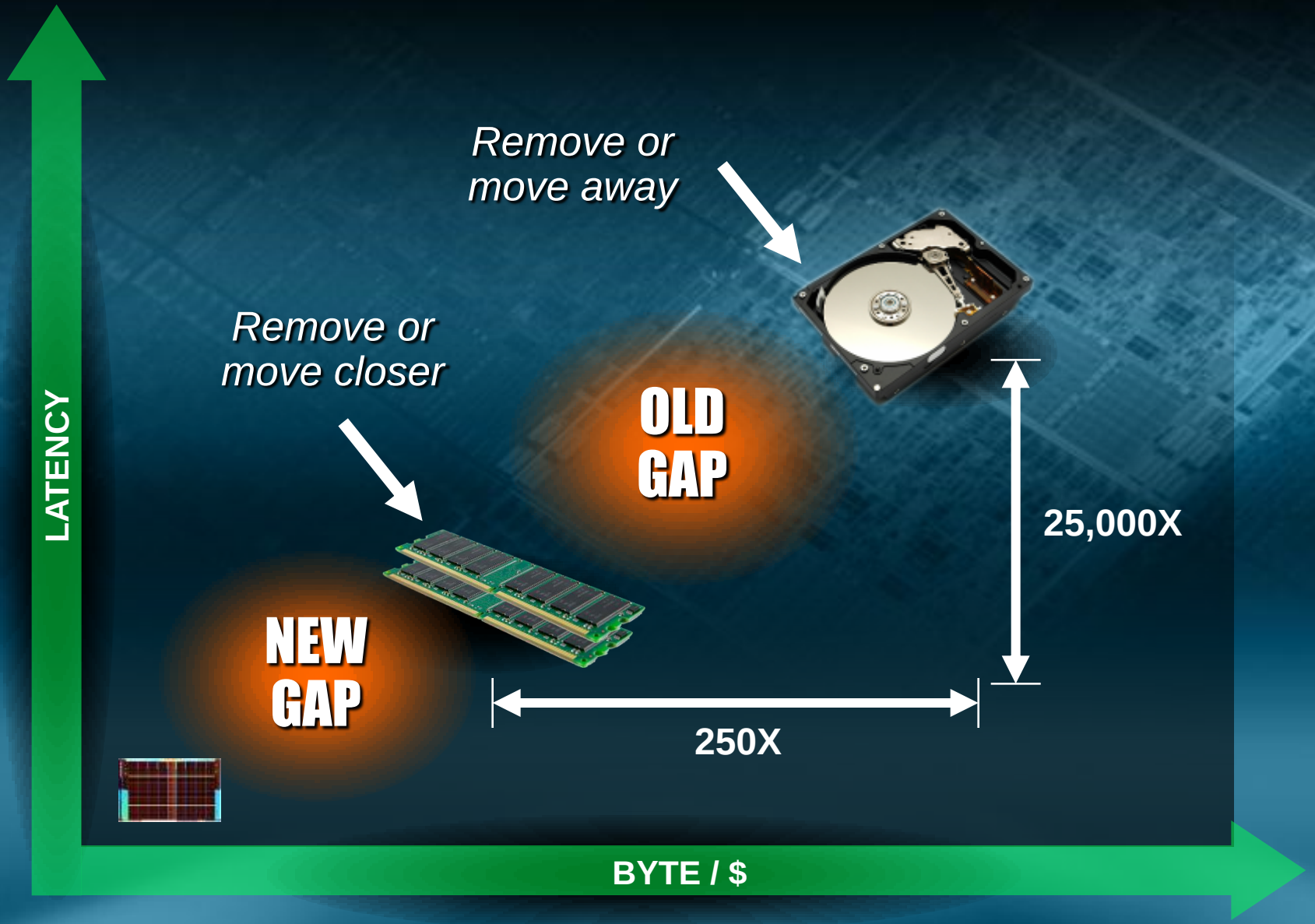


Elapsed Time	45.700667, s
Disk Busy Time	41.056997, s
Average Data Rate	1.37389, MB/s

89% BUSY

intel

Memory Gaps – Old and New

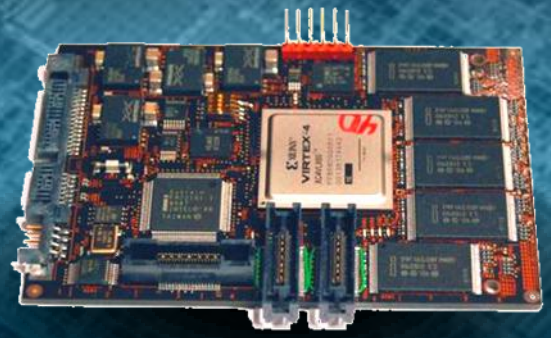


Solid State Drive Performance Variability

Relative IOPS



Intel SATA SSD Prototype



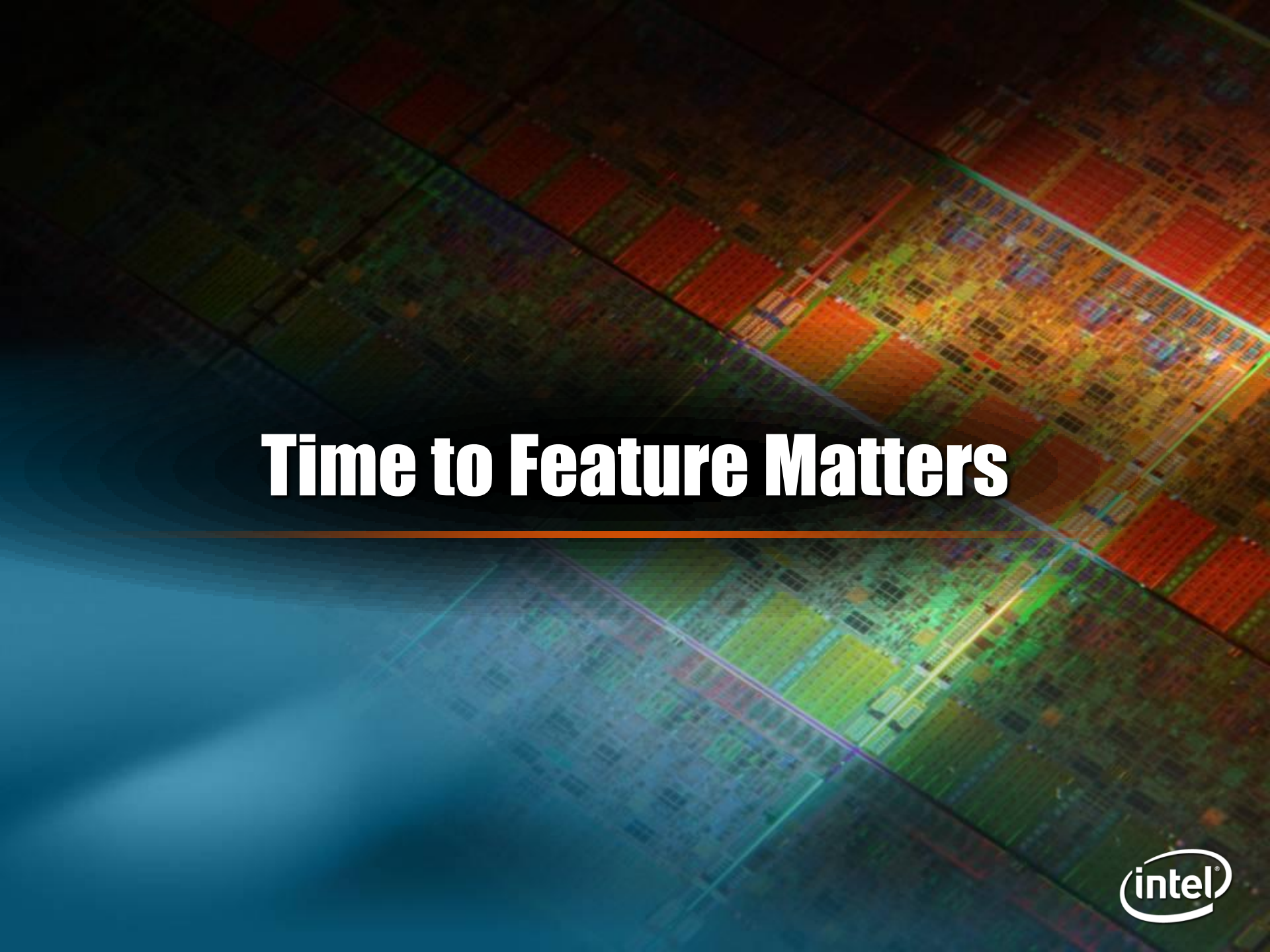
1/10th The Power*

> 10X Performance*

1,000X More Durable*

* Compared to HDDs

intel

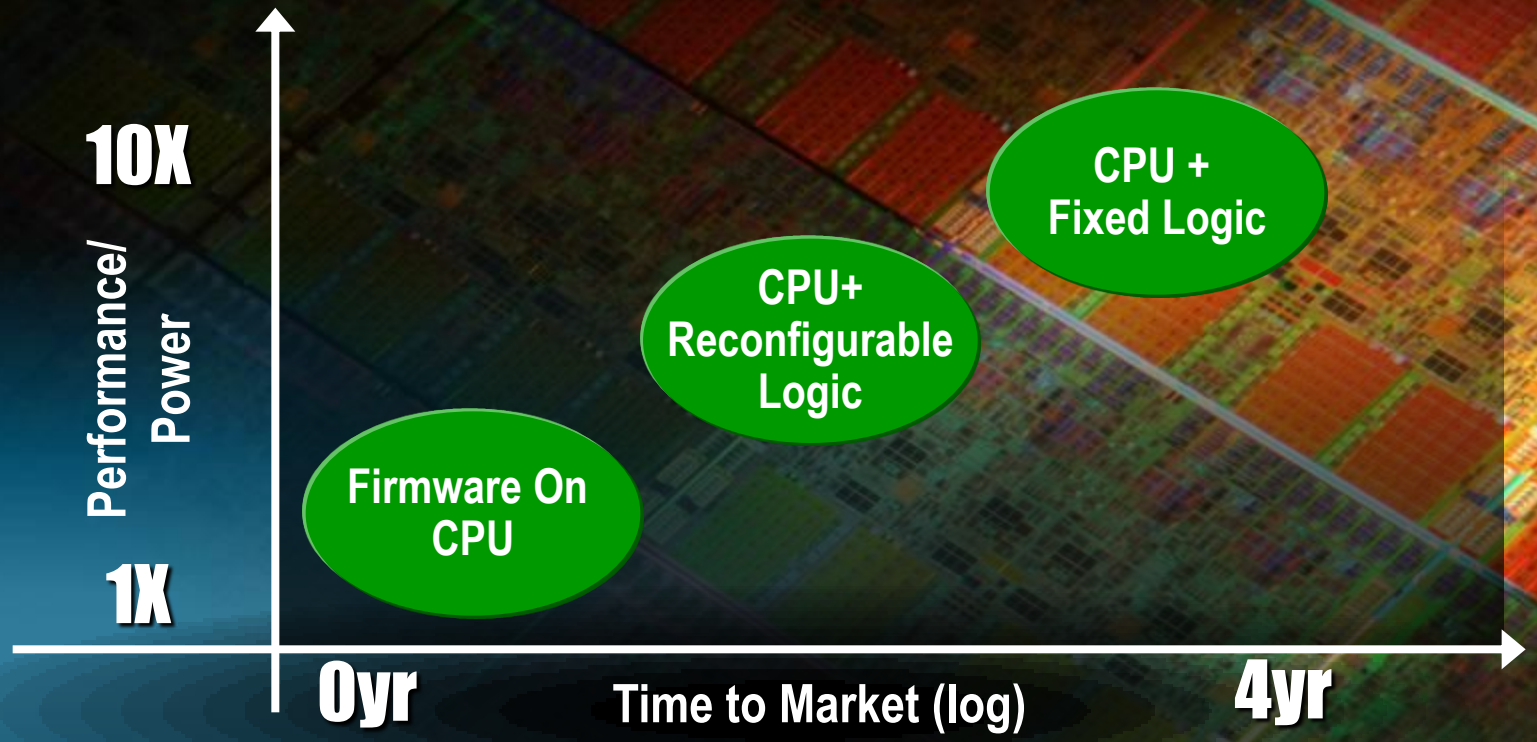

Time to Feature Matters
intel

R&D Pipeline is Very, Very Long

for Mainstream Processors



Feature Implementation Choices

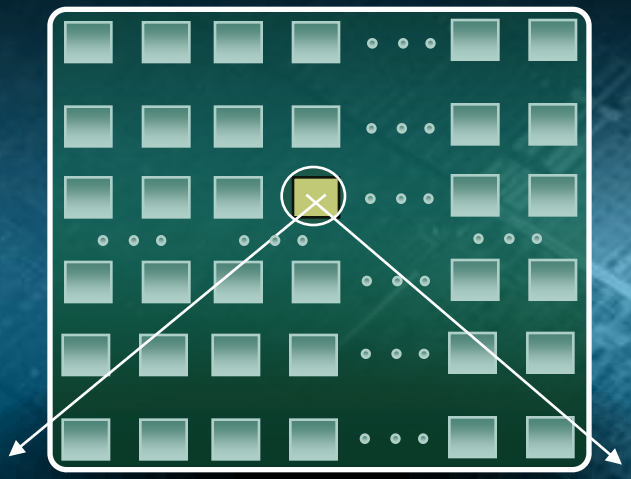


Faster Time to Market
Late Binding of New ISA Features
Quickly Adapting to Changing World

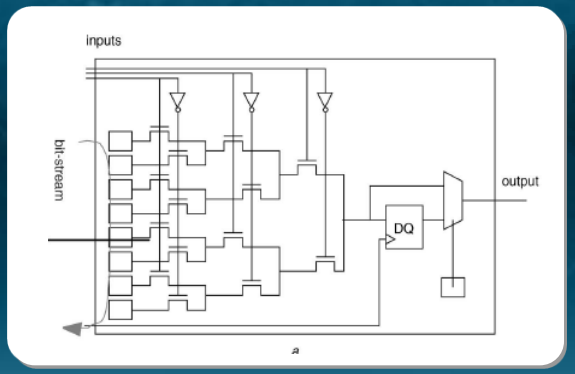
intel

Reconfigurable Logic (RL) Unit

Fine-Grain vs. Coarse-Grain



Greatest Flexibility
Lowest Area Efficiency



Array of Lookup Tables
(i.e. FPGA)

A Hybrid PE Design
Most Suitable

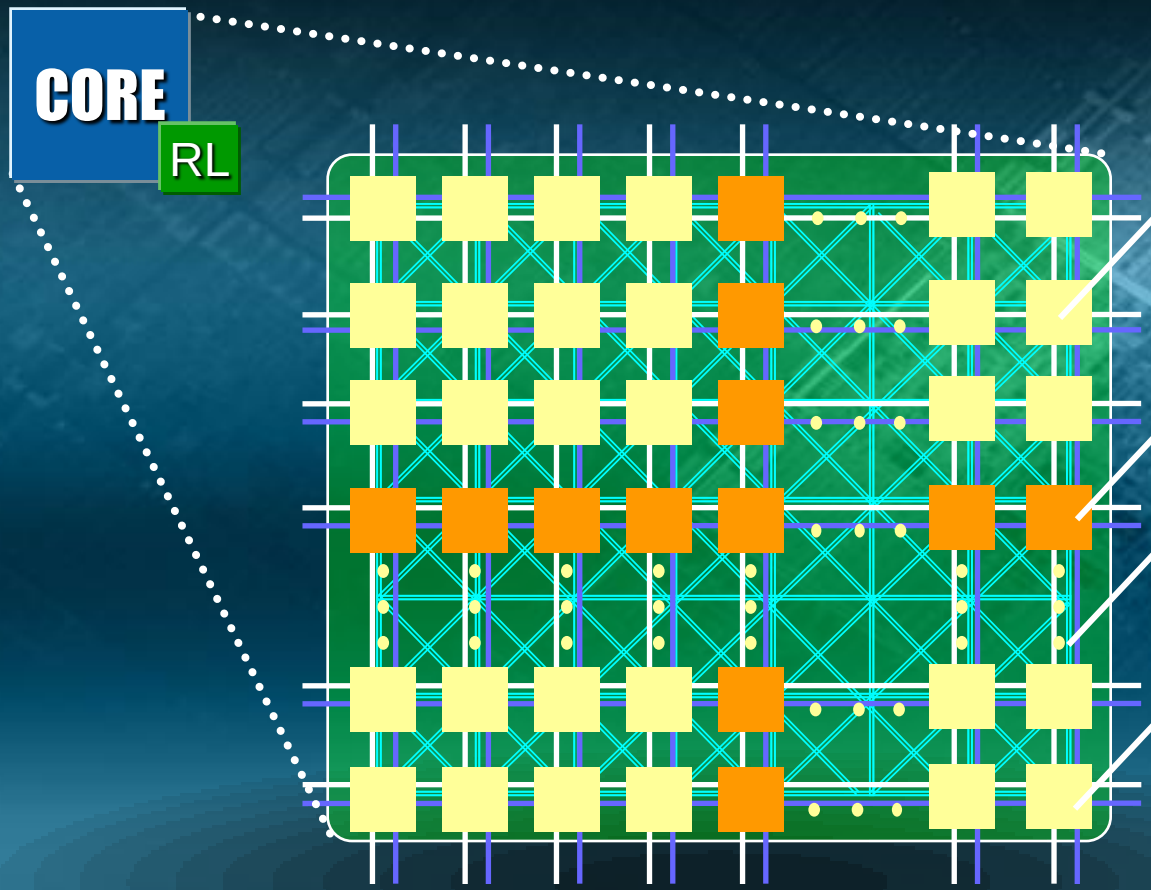
5X Area and Freq Loss

Reduced Flexibility
Higher Area Efficiency

Array of 8-bit ALUs
(e.g. Matrix)

RL can provide a 2-4X benefit over SW
Multiple instructions can be mapped to one RL

Reconfigurable Array of PEs



Regular PEs
(Configurable Logic)

Control PEs
(Configurable Logic)

PE Fabric
(Interconnect Between PEs)

PE I/O
(Data input and output)

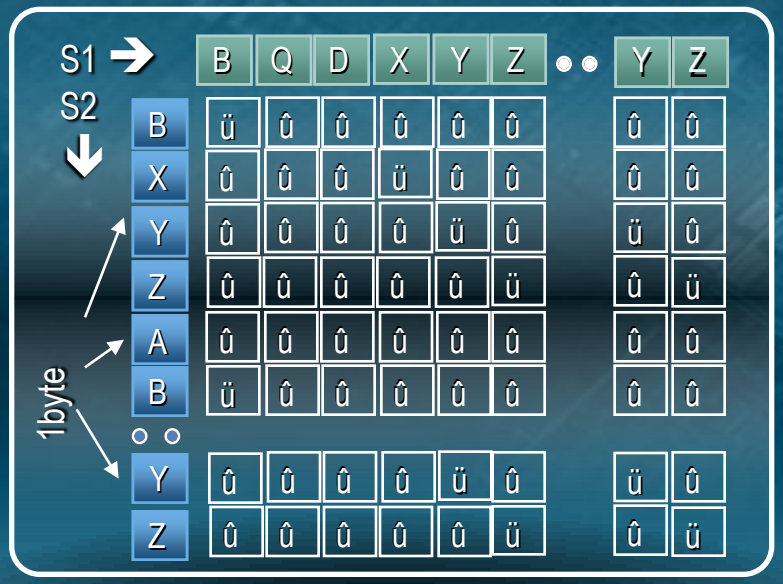
intel

Two Specialization Examples
(Integer/Text)

String Matching

String Processing

XML processing

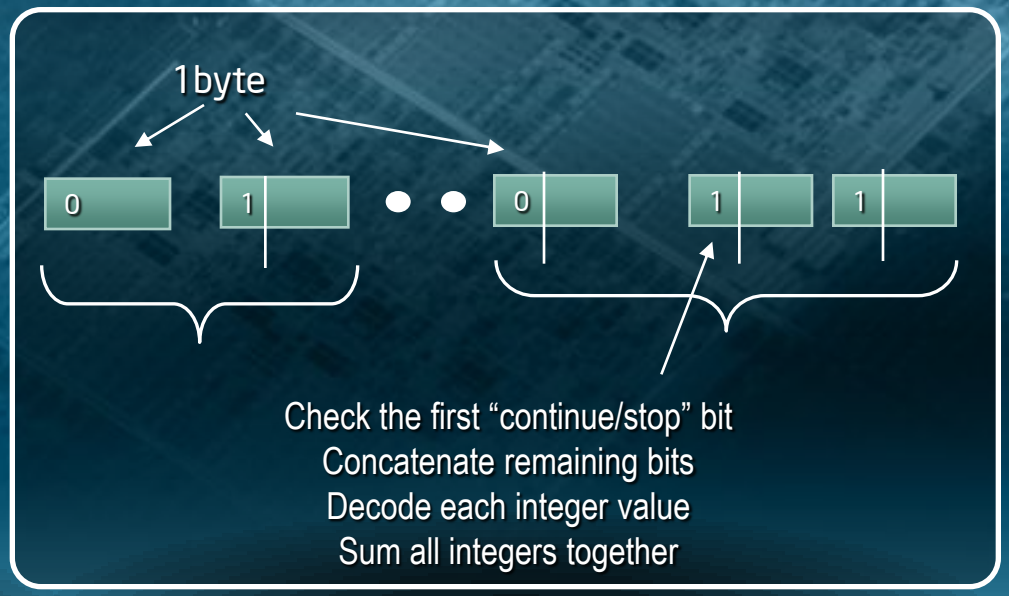


Exhaustive Byte by Byte comparisons inefficient with traditional IA ISA*

Variable Length Decoding

Integer processing

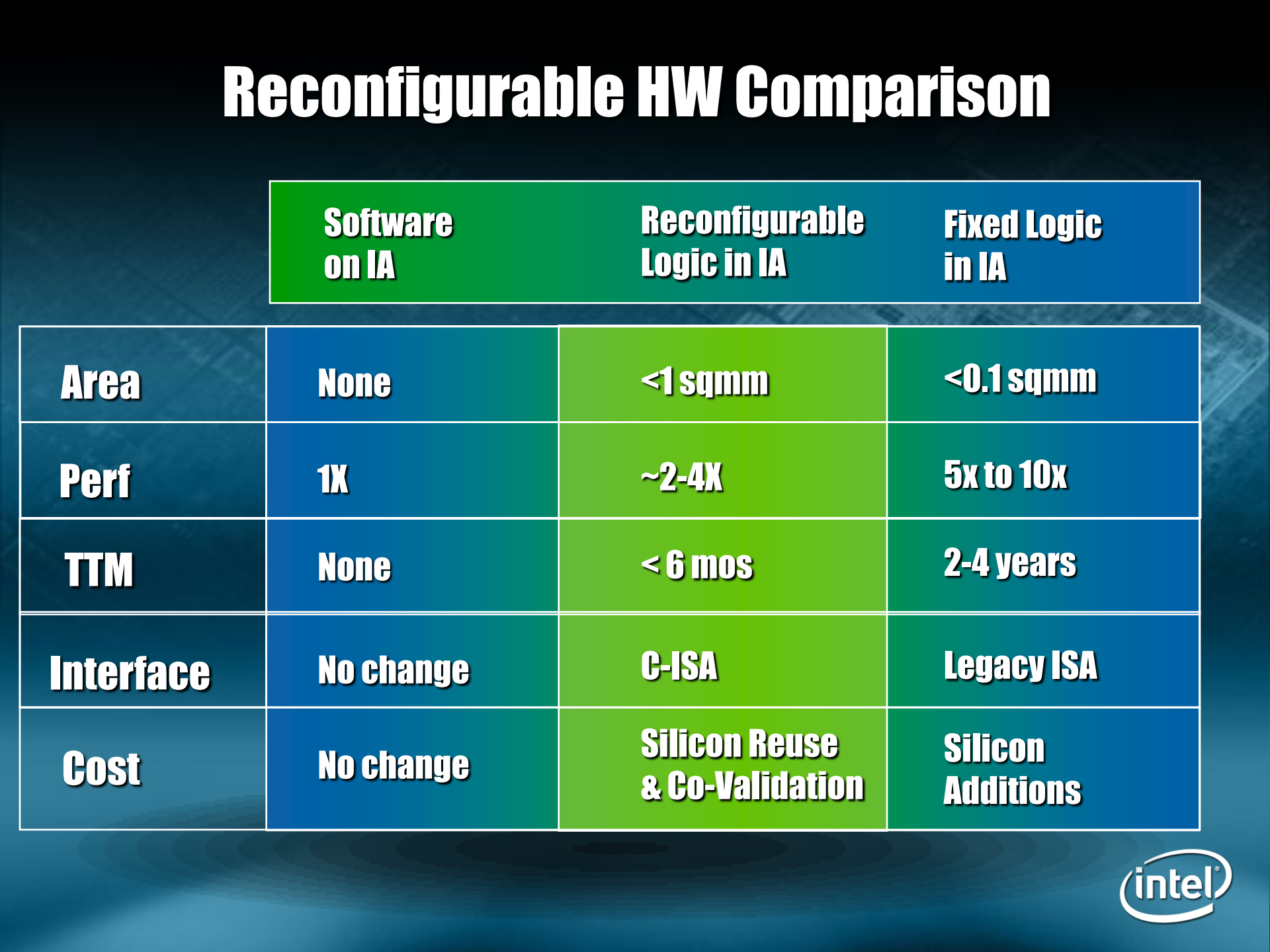
Search Engines

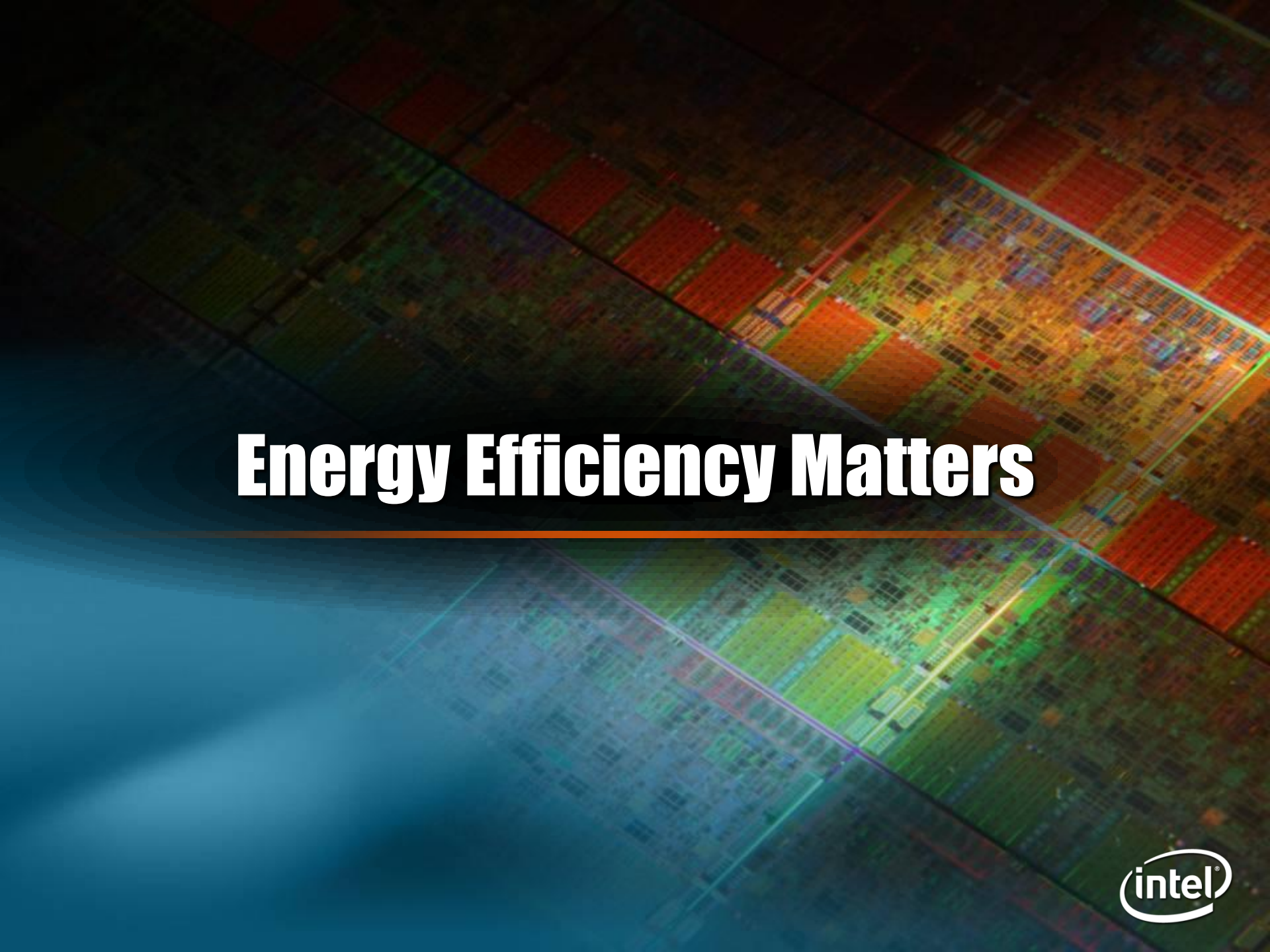


Byte-wise operation
Shifts based on **Bit-based** Control
Inefficient with traditional IA ISA

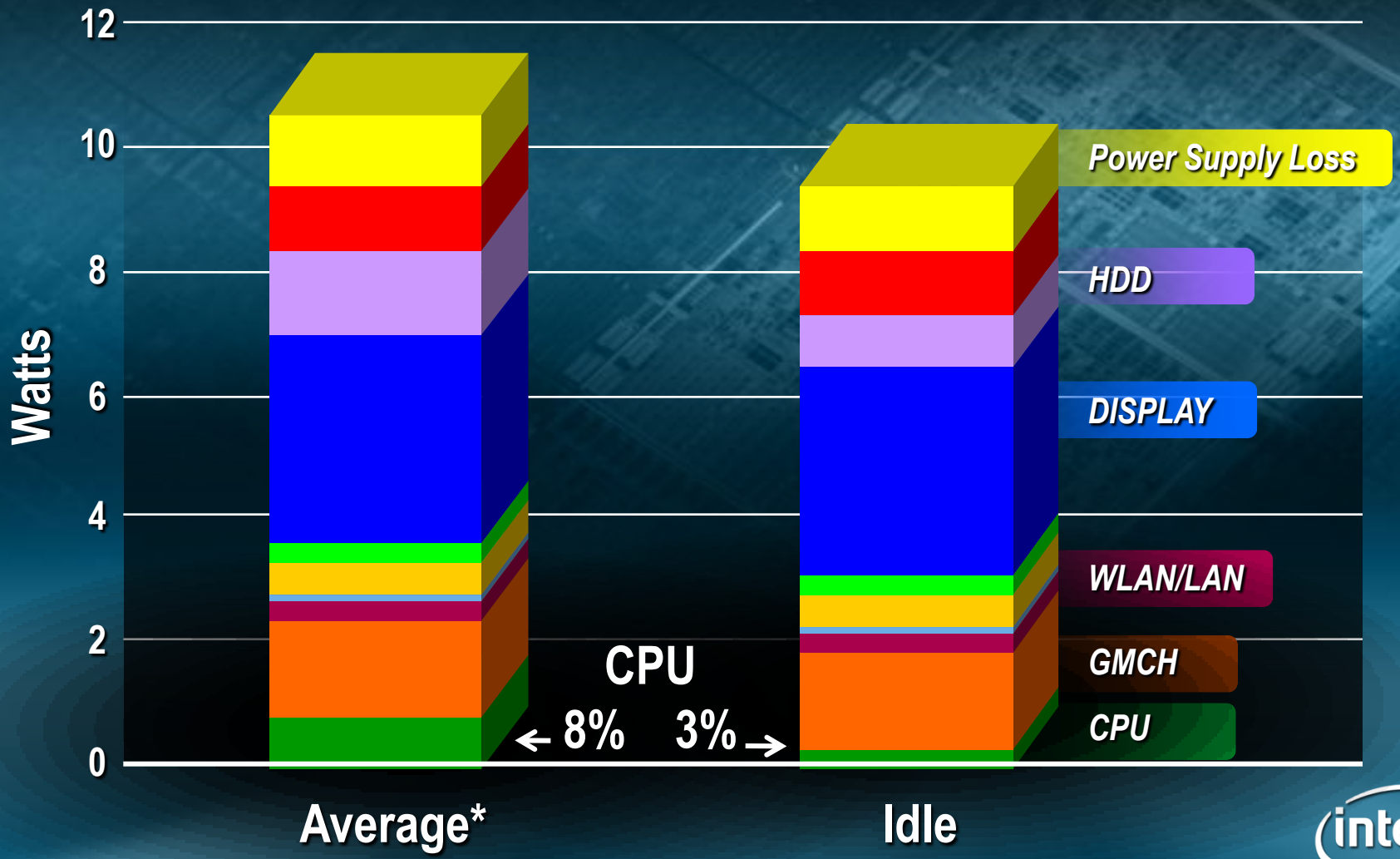
Reconfigurable HW Comparison

	Software on IA	Reconfigurable Logic in IA	Fixed Logic in IA
Area	None	<1 sqmm	<0.1 sqmm
Perf	1X	~2-4X	5x to 10x
TTM	None	< 6 mos	2-4 years
Interface	No change	C-ISA	Legacy ISA
Cost	No change	Silicon Reuse & Co-Validation	Silicon Additions



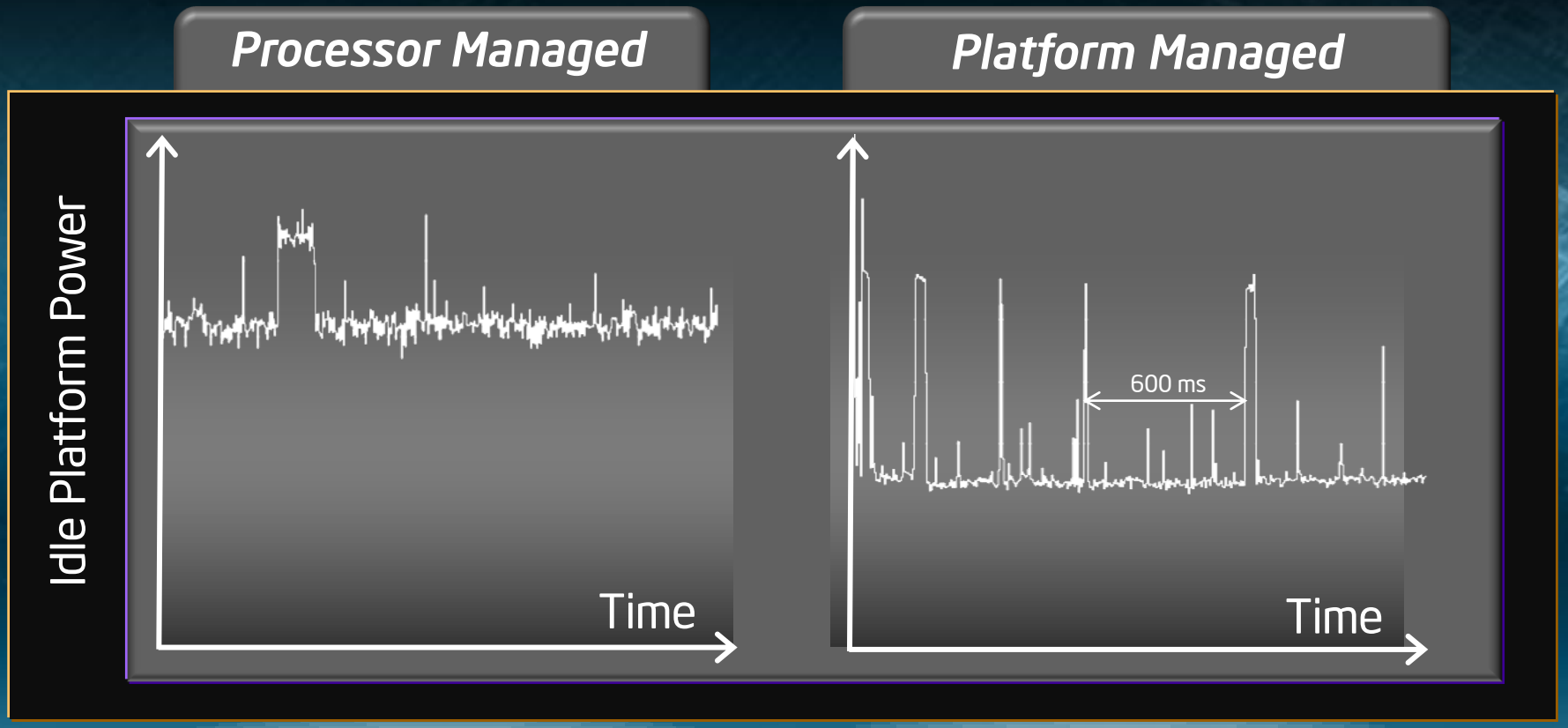

Energy Efficiency Matters
intel

Energy Efficiency: Average Vs. Idle



* Mobile Mark 2005

50% Reduction in Idle Power



- Periodic, polled activity
- Frequent, asynchronous events
- OS changes power state

- Tickless, event-driven OS
- Grouped & aligned activity
- HW changes power state

Holistic Approach to Energy Efficiency

Platform Managed Power

Operating Systems and VMMs
Well-behaved software
OS/VMM-guided PM policies

Manageability
Scale power management from
a system to the data center

Telemetry
Enhanced visibility for
platform-level policies
(e.g. power, temperature)



Core Logic
Platform-level power, performance and
thermal management utilizing a
rich set of fine-grain techniques

Interconnects and Peripherals
Well-behaved, power-efficient devices
and interconnects

Power Delivery and Cooling
Maximum efficiency under all loads


Immersiveness Matters
intel

Hints of a Revolution

Google Earth



iPhone



Second Life



Nintendo Wii



Medical Imaging



Robot Cars



Virtual Worlds Compute Requirements

Client

Application	% CPU Utilization	% GPU Utilization (nVidia G80)
2D Websites	20	0-1
Google Maps	60	3-5
Google Earth	50	10-15
Google 3D Warehouse	50	15-20
Second Life (SL)	70	35-75

SL requires 3X more CPU processing and 10-100X more GPU processing than 2D websites

SL client spends 65%+ of CPU time in compute-intensive components

SL needs at least 20x GPU processing compared to 2D

Server

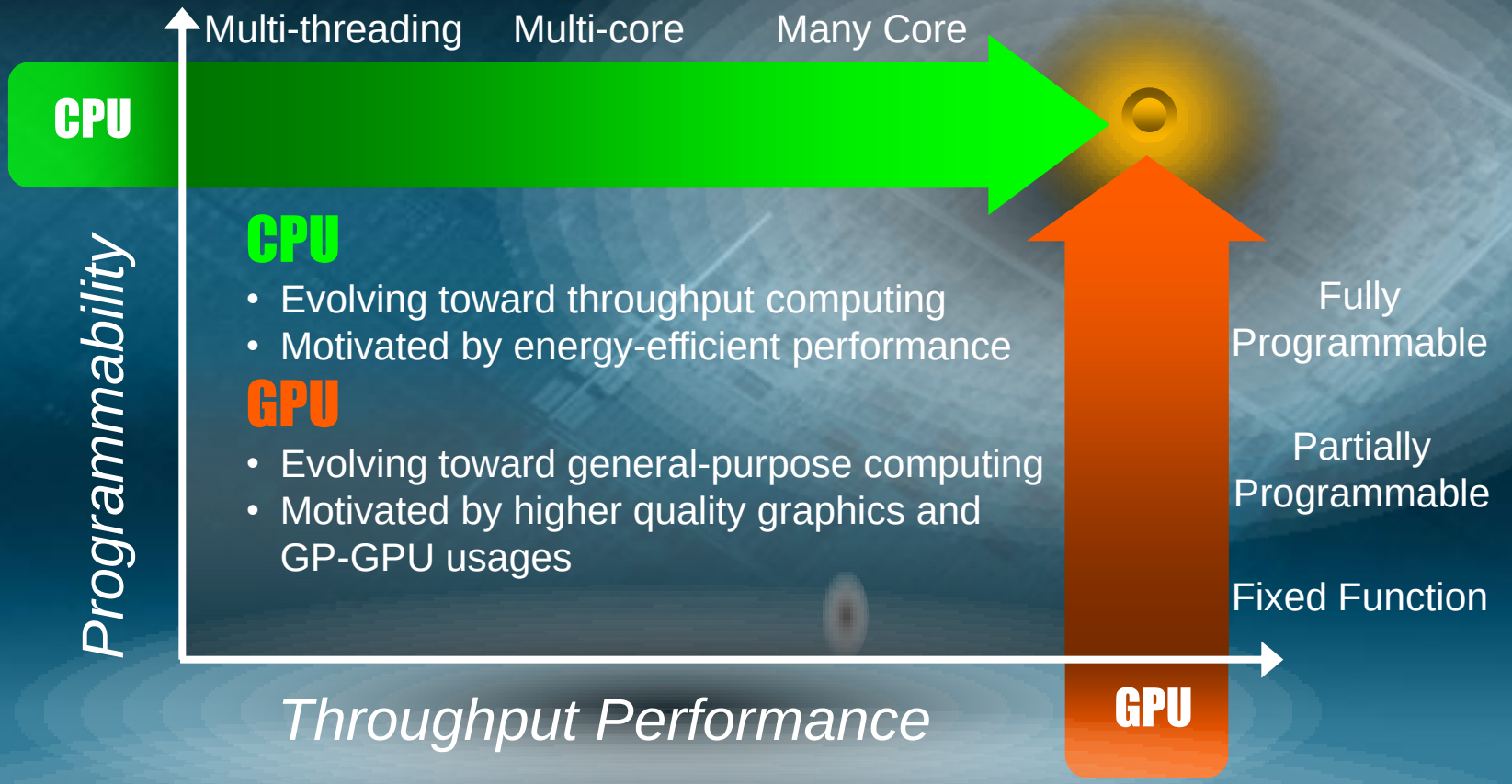
Type	Software	Maximum Client/Server
MMOG	Eve Online	34420
	WoW	2500
Virtual Worlds	Second Life	40-60

SL server-side spends 75%+ time in compute-intensive components

SL requires 10-100x more computation per client than MMORGS



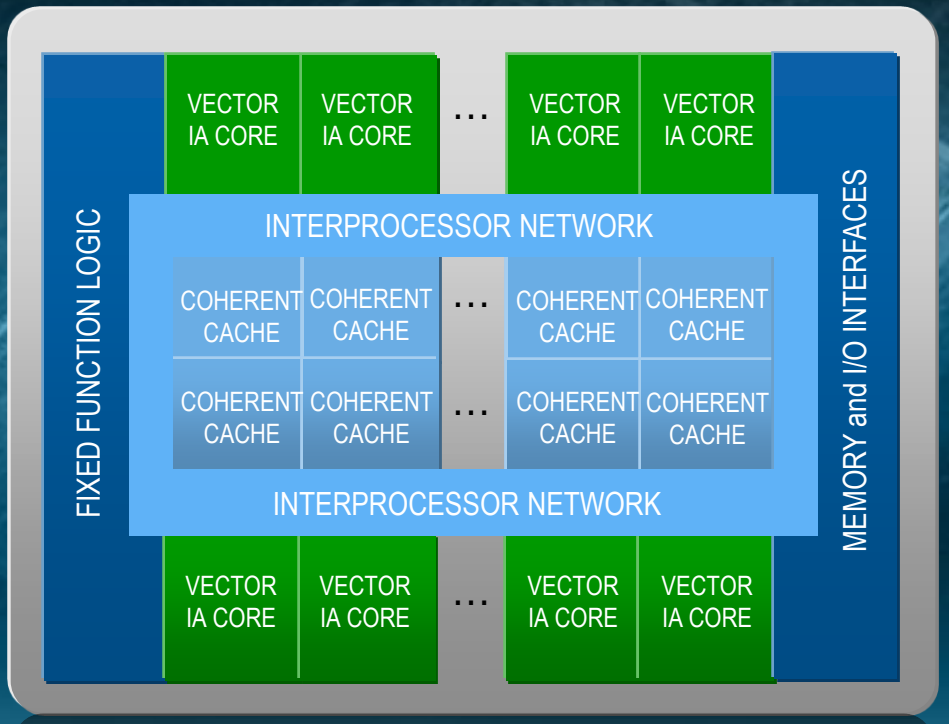
When Architectures Collide



An Architecture for Immersive Computing

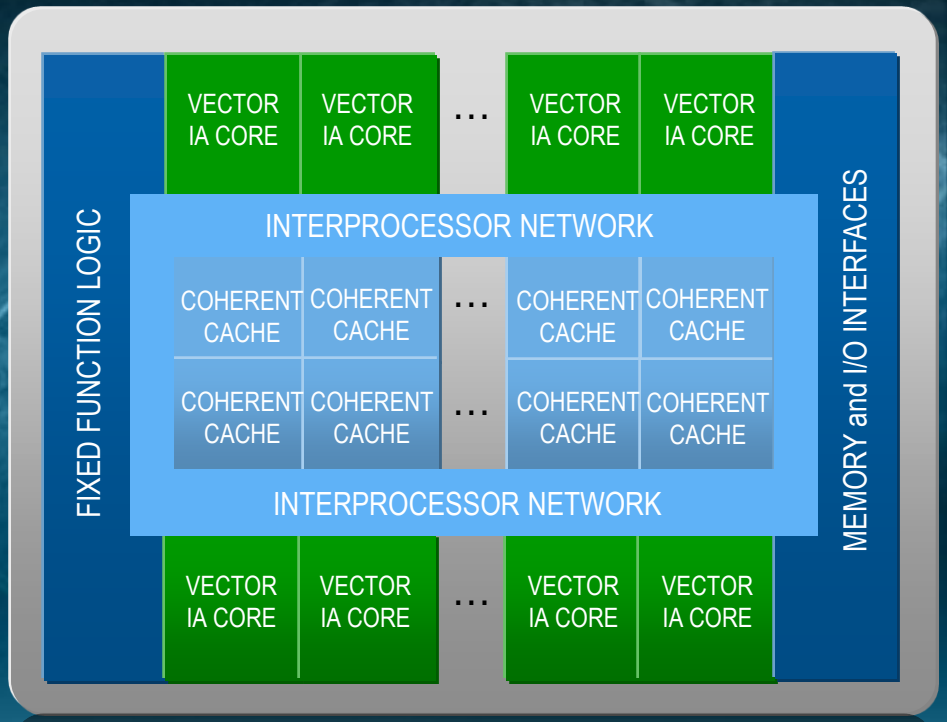
Intel's Forthcoming Larrabee Processor

- Familiar programming model
- Support for irregular, shared data
- Automatic data movement
- Algorithmic vs. instruction efficiency



Let's Take a Moment to Clarify

- LRB is NOT a Ray-Tracing Machine
- LRB is also NOT a Raster Machine
- LRB does both very well in software
- LRB does other things quite well, too*
- Significant LRB disclosures at
  - SIGGRAPH '08
  - HOTCHIPS '08
  - VLDB '08



*

PARSEC Benchmark Suite

Released January 29, 2008

Princeton University · intel · UPMC University of Pittsburgh Medical Center

Challenges Going Forward



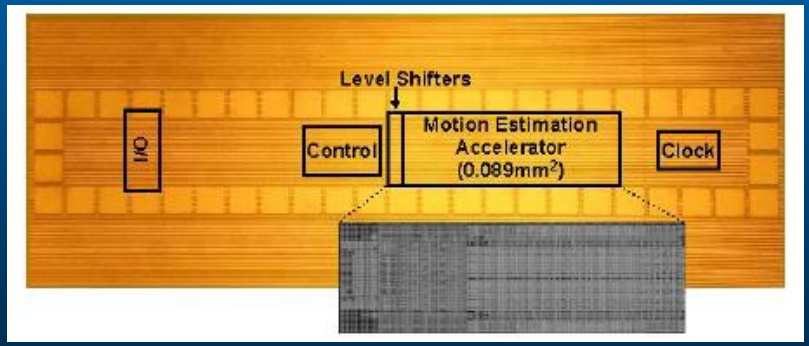
Optimal Core Size

13mm, 100W, 48MB Cache,
4B Transistors, in 22nm



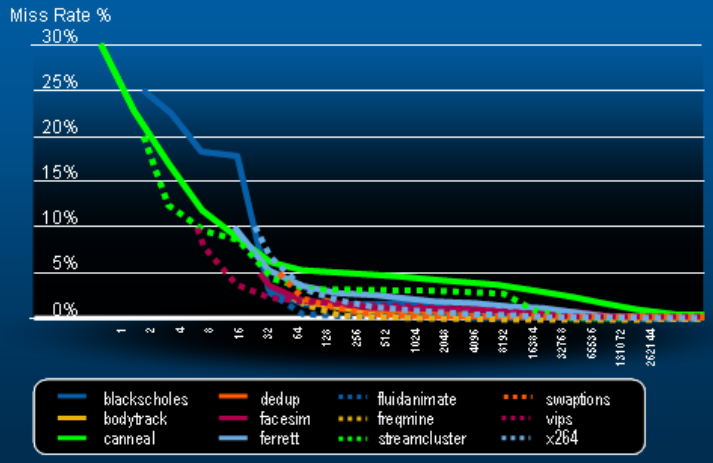
Sub-Threshold Logic

320mV 56µW 411GOPS/W Ultra-Low Voltage
Motion Estimation Accelerator



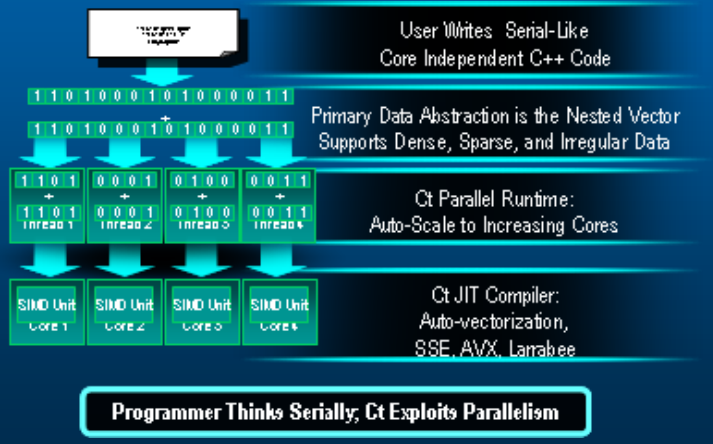
Memory Bandwidth

Enabling Large Capacity LLC's



Programming Languages

All Programs Become Parallel Programs



Programmer Thinks Serially; Ct Exploits Parallelism

What Really Matters?

Response Time
Time to Feature
Platform Energy Efficiency
Immersive Experience



Xie Xie

