The Future Is Heterogeneous Computing

Mike Houston

Principal Architect, Accelerated Parallel Processing
Advanced Micro Devices

October 27th, 2010

Workload Example: *Changing Consumer Behavior*

20 hours
of video
uploaded to YouTube
every minute

Approximately
9 billion
video files owned are
high-definition

50 million +
digital media files
added to personal content libraries
every day

1000
images
are uploaded to Facebook
every second

Challenges for Next Generation Systems

- The Power Wall
  - Even more broadly constraining in the future!
- Complexity Management – *HW and SW*
  - Principles for managing exponential growth
- Parallelism, Programmability and Efficiency
  - Optimized SW for System-level Solutions
- System balance
  - Memory Technologies and System Design
  - Interconnect Design

The Power Wall

- Easy prediction: *Power will continue to be the #1 design constraint for Computer Systems design.*
- Why?
  - *V_{min} will not continue tracking Moore's Law*
  - *Integration of system-level components consume chip power*
    - *A well utilized 100GB/sec DDR memory interface consumes ~15W for the I/O alone!*
  - *2nd Order Effects of Power*
    - *Thermal, packaging & cooling (node-level & datacenter-level)*
    - *Electrical stability in the face of rising variablity*
  - *Thermal Design Points (TDPs) in all market segments continue to drop*
  - *Lightly loaded and idle power characteristics are key parameters in the Operational Expense (OpEx) equation*
  - *Percent of total world energy consumed by computing devices continues to grow year-on-year*

Optimized SW for System-level Solutions

- Long history of SW optimizations for HW “characteristics”
  - Optimizing compilers
  - Cache / TLB blocking
  - Multi-processor coordination: communication & synchronization
  - Non-uniform memory characteristics: Process and memory affinity
- Scarcity/Abundance principle favors increased use of Abstractions
  - Abstraction leads to Increased productivity but costs performance
  - Still allow experts burrow down into lower level “on the metal” details
- *System-level Integration Era will demand even more*
  - Many Core: user mode and/or managed runtime scheduling?
  - Heterogeneous Many Core: capability aware scheduling?
- *SW productivity versus optimization dichotomy*
  - Exposed HW leads to better performance but requires a “platform characteristics aware programming model”

The Memory Wall – *getting thicker*

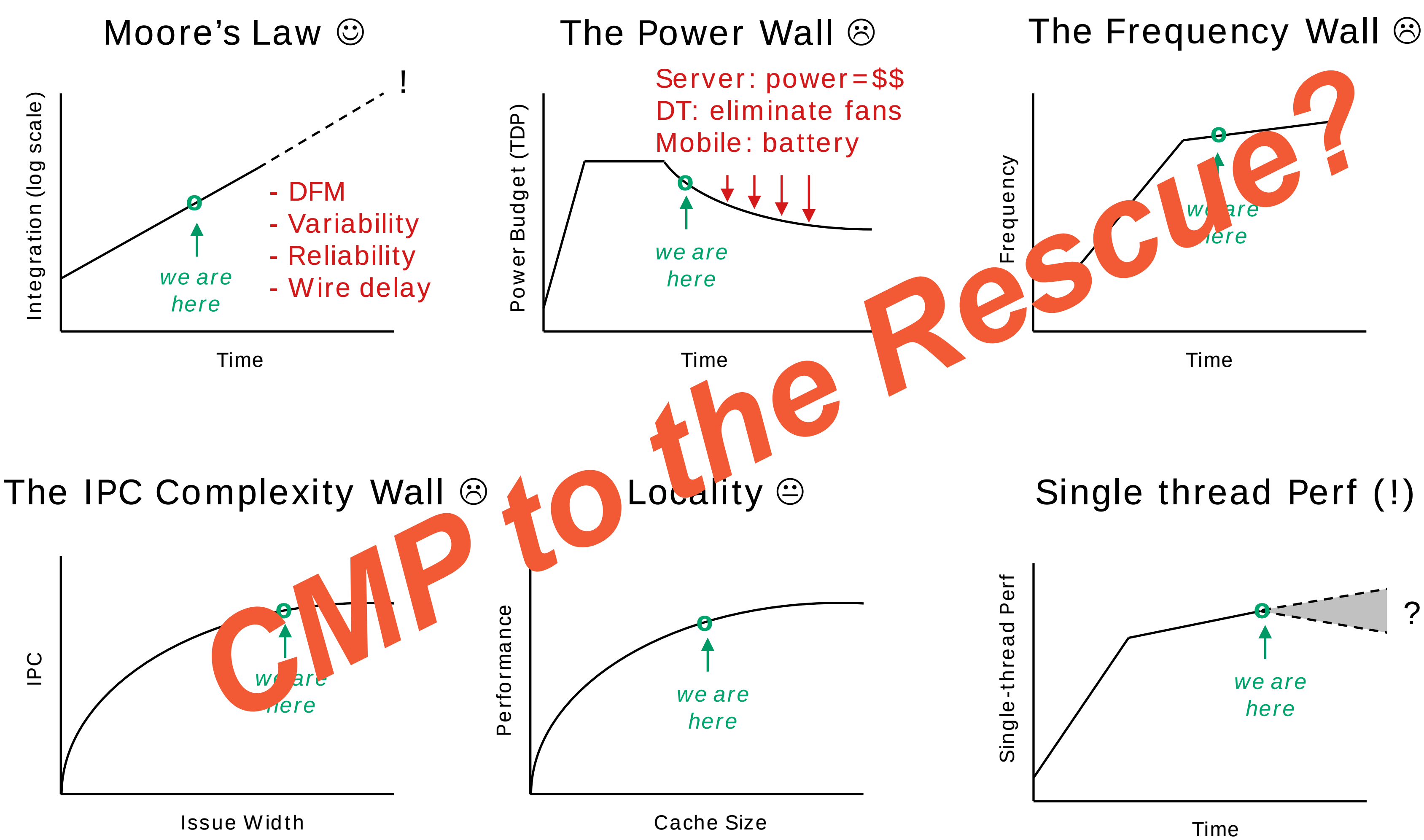
There has always been a Critical Balance between
Data Availability and *Processing*

Situation	When?	Implication	Industry Solutions	
DRAM vs CPU Cycle Time Gap	Early 1990s	Memory wait time dominates computing	**Non-blocking caches** **O-o-O Machines**	☺
SW Productivity Crisis Object oriented languages; Managed runtime environments	Mid 1990s	Larger working sets More diverse data types	**Larger Caches** **Cache Hierarchies** **Elaborate prefetch**	😐
Single Thread → CMP Focus	2005 and beyond	Multiple working sets! Virtual Machines! More memory accesses	**Huge Caches** **Multiple Memory Controllers** **Extreme PHYs**	😐 ↓ ☹
New & Emerging Abstractions Browser-based Runtimes Image/Video as basic data types Throughput-based designs	2009 and beyond	Even larger working sets Larger data types	***Accelerated Parallel Processing*** ***Chip Stacking***	*TBD*

Interconnect Challenges

- Coherence domain – *knowing when to stop*
  - Interesting implications for on-chip interconnect networks
- Industry Mantra: *"Never bet against Ethernet"*
  - But, current Ethernet not well suited for lossless transmission
  - Troublesome for storage, messaging and more
- The more subtle and trickier problems
  - Adaptive routing, congestion management, QOS, End-to-end characteristics, and more
- Data centers of tomorrow are going to take great interest in this area

Single-thread Performance

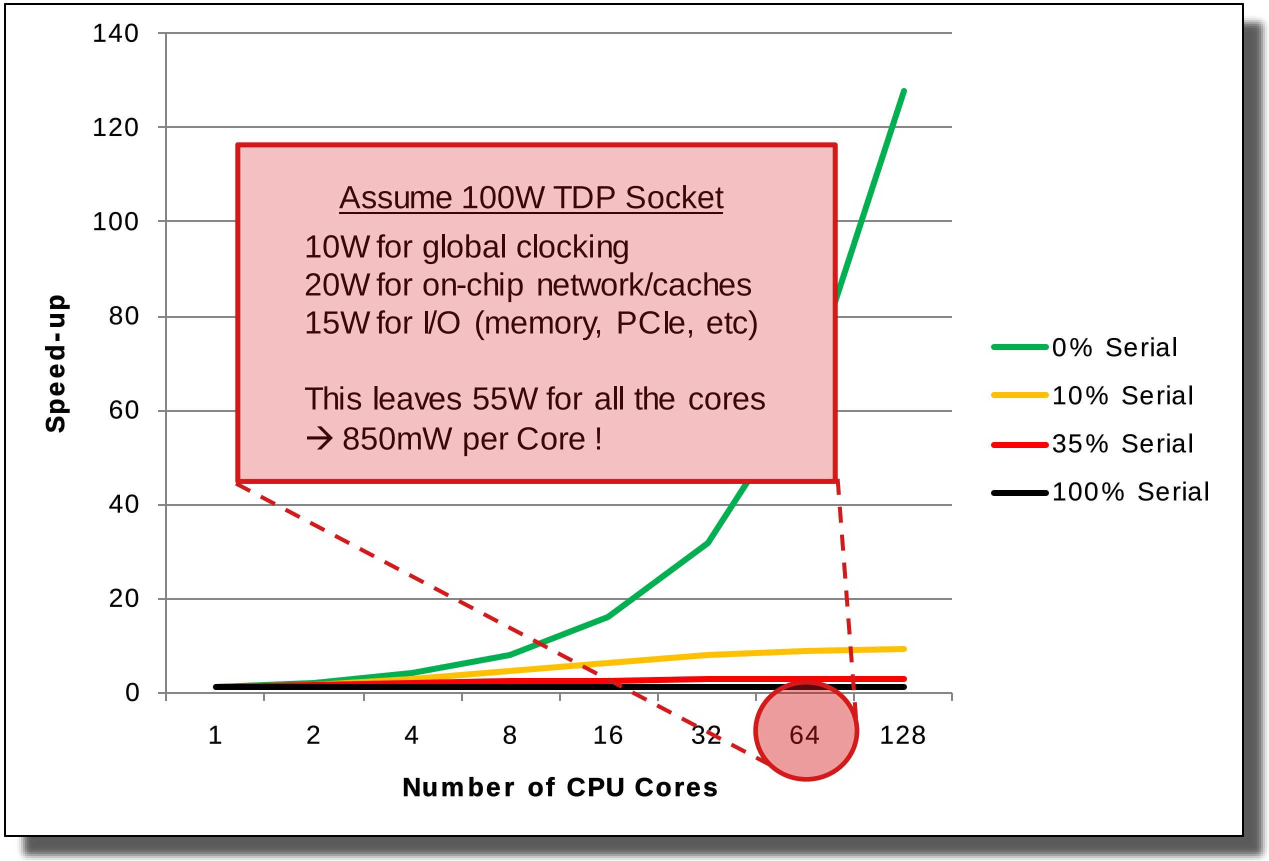


Parallel Programs and Amdahl's Law

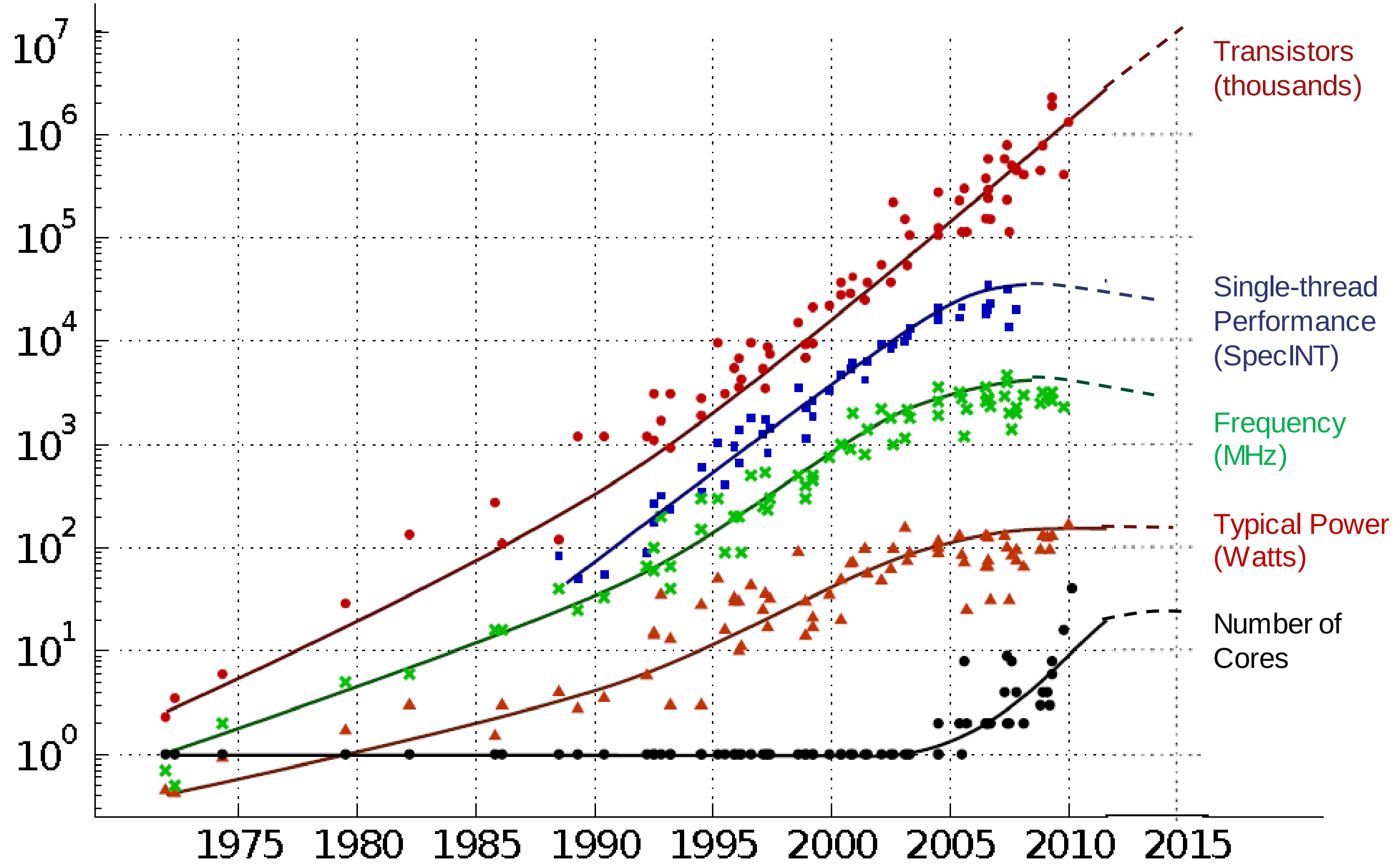
$$\text{Speed-up} = \frac{1}{S_W + (1 - S_W) / N}$$

S_W: % Serial Work

N: Number of processors



35 Years of Microprocessor Trend Data



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

The Power Wall – *Again!*

- *Escalating multi-core designs will crash into the power wall just like single cores did due to escalating frequency*

- Why?
  - In order to maintain a reasonable balance, core additions must be accompanied by increases in other resources that consume power (on-chip network, caches, memory and I/O BW, ...)
    - Spiral upwards effect on power
  - The use of multiple cores forces each core to actually **slow down**
    - At some point, the power limits will not even *allow* you to activate all of the cores at the same time
  - Small, low-power cores tend to be very weak on single-threaded general purpose workloads
    - Customer value proposition will continue to demand excellent performance on general purpose workloads
    - The transition to compelling general purpose parallel workloads will not be a fast one

What about Throughput Computing?

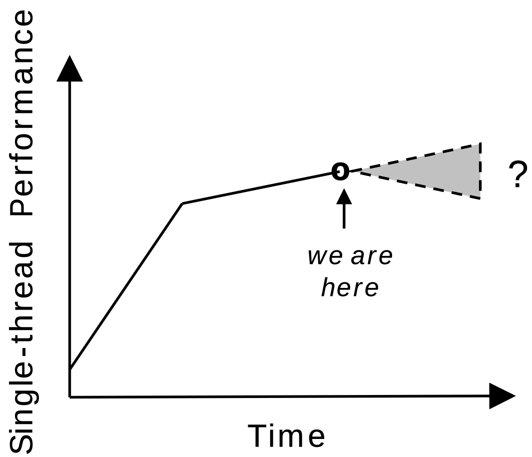
- Works around Amdahl's law by focusing on *throughput* of multiple independent tasks
  - Servers: Transaction Processing; Web Clicks; Search Queries
  - Clients: Graphics; Multimedia; Sensory Inputs *(future)*
  - HPC: Data-level parallelism
- New bottlenecks start to appear
  - As some point, the OS itself becomes the "serial component"
    - User mode scheduling and task-stealing runtimes
  - Memory BW – Goal is to saturate the pipeline to memory
    - Large number of outstanding references
    - Large number of active and/or standby threads
  - Power – Overall utilization goes up, so does power consumption
    - Still the #1 constraint in modern computer design

Three Eras of Processor Performance

Single-Core Era

Enabled by:
- ✓ Moore's Law
- ✓ Voltage Scaling
- ✓ MicroArchitecture

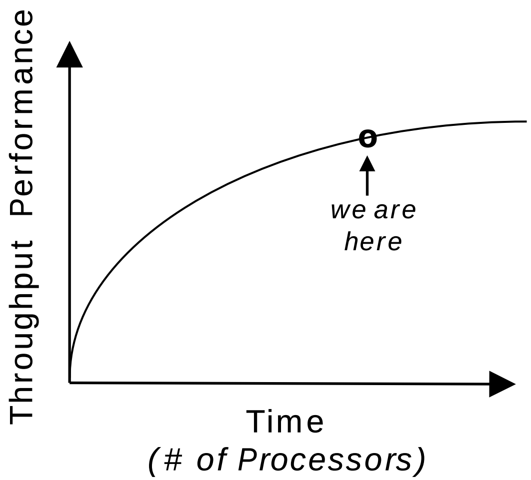
Constrained by:
- ✘ Power
- ✘ Complexity



Multi-Core Era

Enabled by:
- ✓ Moore's Law
- ✓ Desire for Throughput
- ✓ 20 years of SMP arch

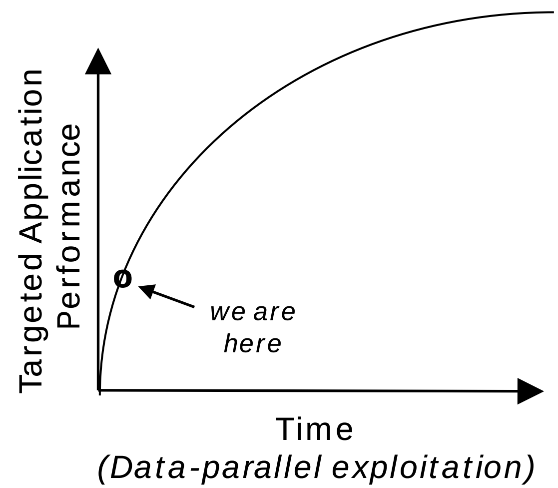
Constrained by:
- ✘ Power
- ✘ Parallel SW availability
- ✘ Scalability



Heterogeneous Systems Era

Enabled by:
- ✓ Moore's Law
- ✓ Abundant data parallelism
- ✓ Power efficient GPUs

Currently constrained by:
- ✘ Programming models
- ✘ Communication overheads

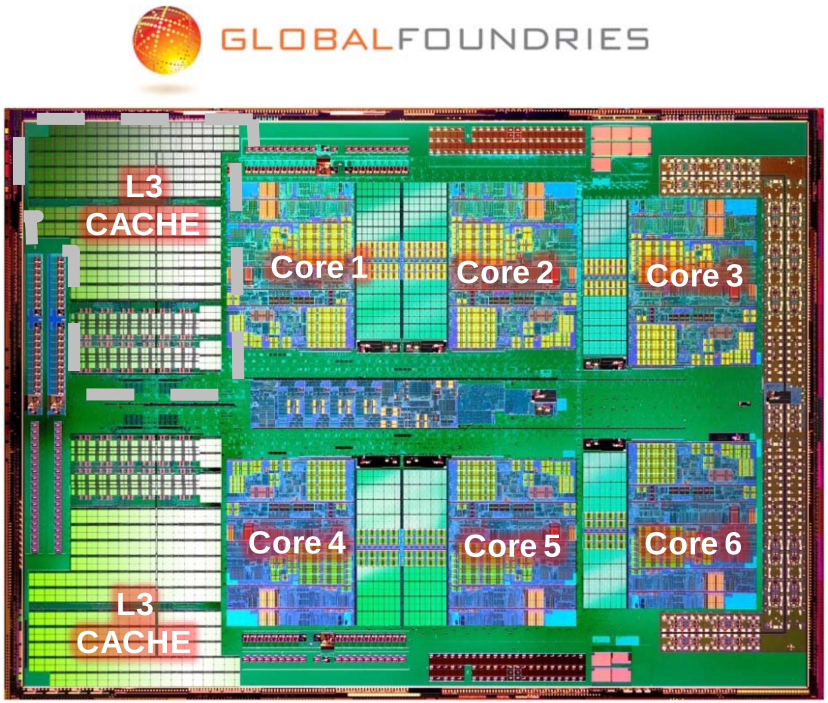


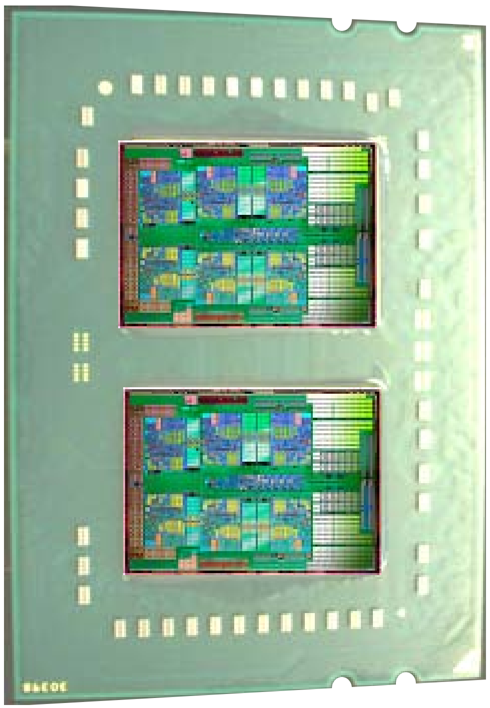
AMD x86 64-bit CMP Evolution

	2003	2005	2007	2008	2009	2010
	AMD Opteron™	Dual-Core AMD Opteron	Quad-Core AMD Opteron	45nm Quad-Core AMD Opteron	Six-Core AMD Opteron	AMD Opteron 6100 Series
Mfg. Process	90nm SOI	90nm SOI	65nm SOI	45nm SOI	45nm SOI	45nm SOI
CPU Core	K8	K8	Greyhound	Greyhound+	Greyhound+	Greyhound+
L2/L3	1MB/0	1MB/0	512kB/2MB	512kB/6MB	512kB/6MB	512kB/12MB
Hyper Transport™ Technology	3x 1.6GT/.s	3x 1.6GT/.s	3x 2GT/s	3x 4.0GT/s	3x 4.8GT/s	4x 6.4GT/s
Memory	2x DDR1 300	2x DDR1 400	2x DDR2 667	2x DDR2 800	2x DDR2 1066	4x DDR3 1333

Max Power Budget Remains Consistent

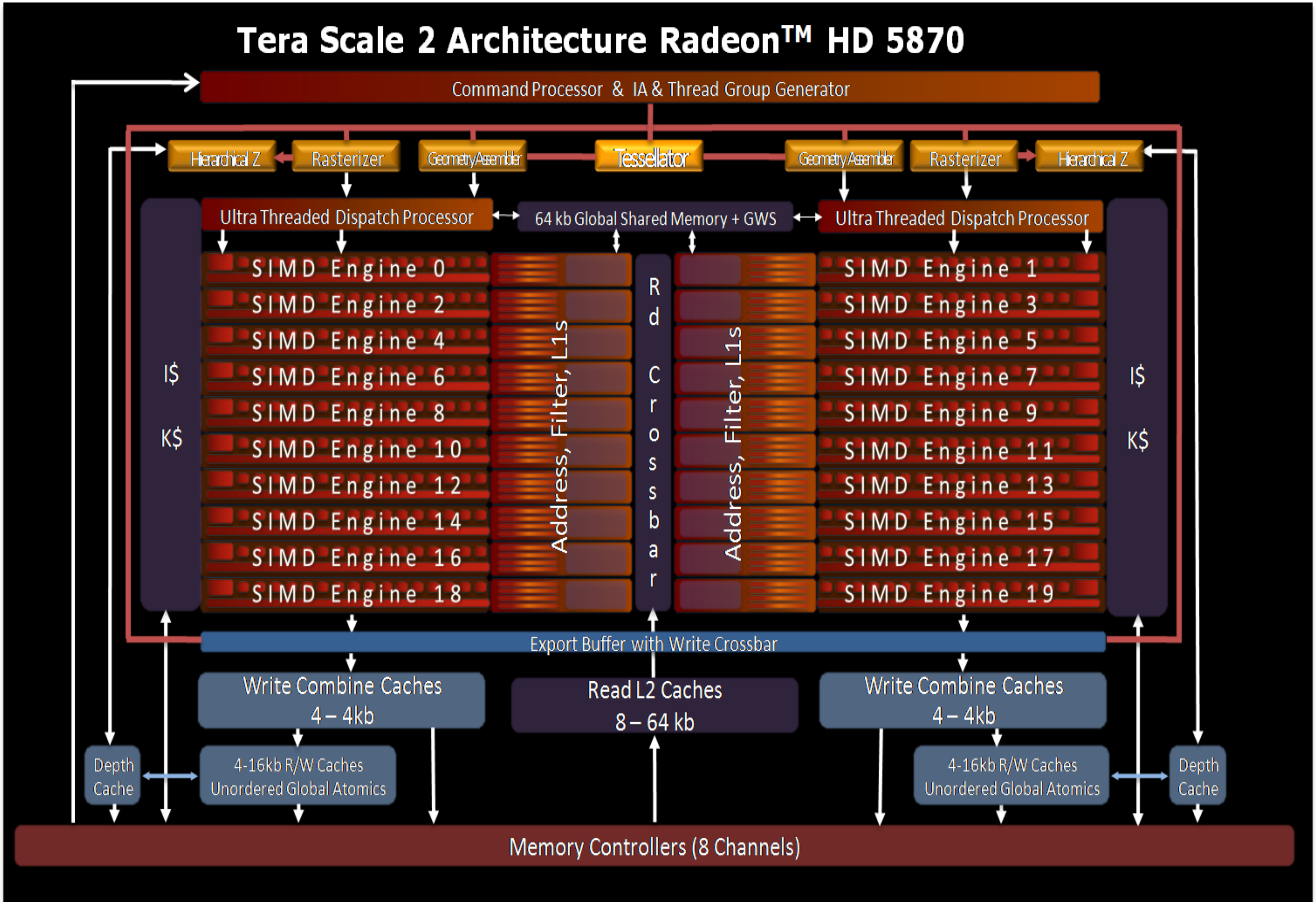
AMD Opteron™ 6100 Series Silicon and Package



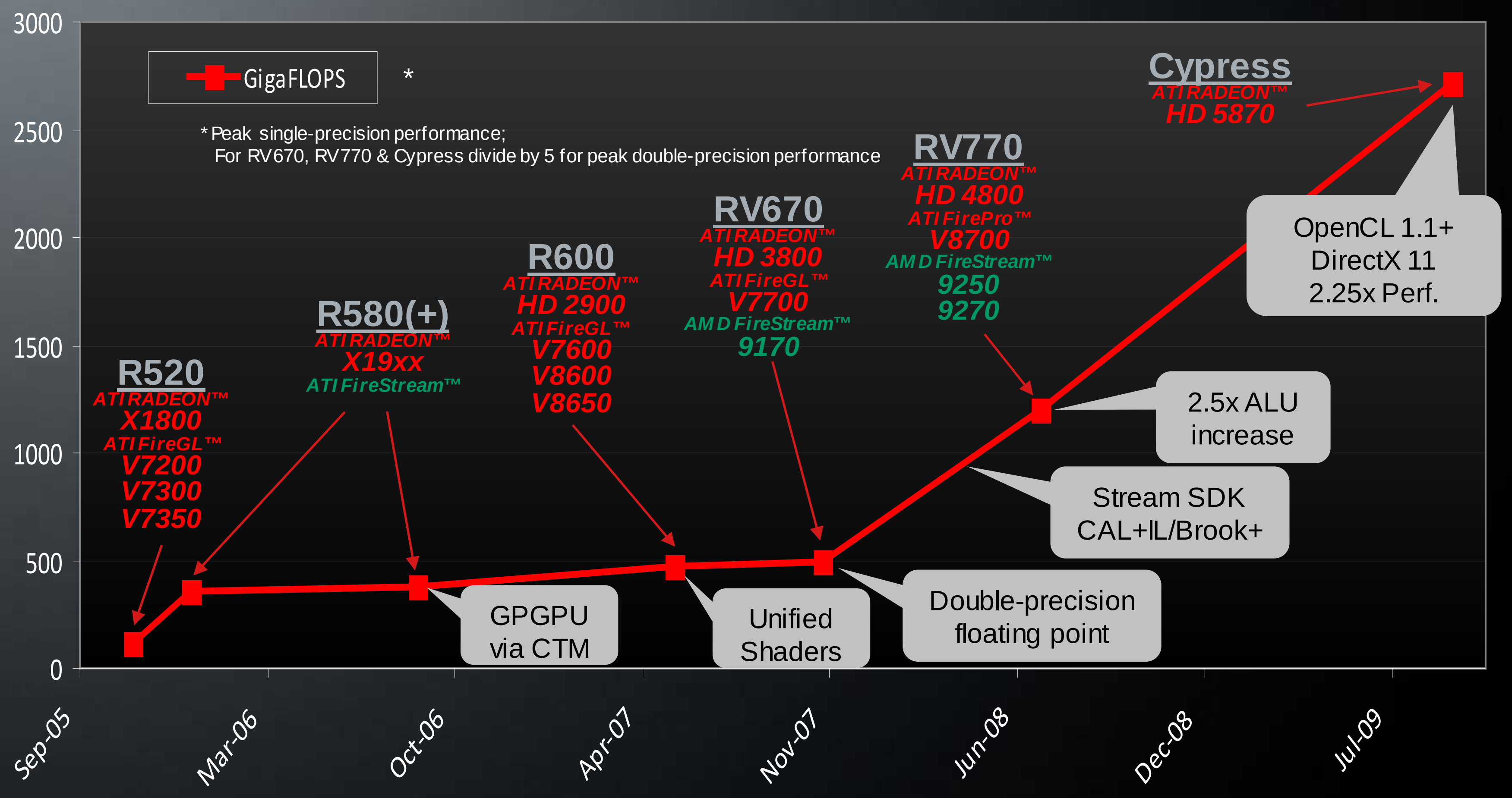


12 AMD64 x86 Cores
18 MB on-chip cache
4 Memory Channels @ 1333 MHz
4 HT Links @ 6.4 GT/sec

AMD Radeon HD5870 GPU Architecture

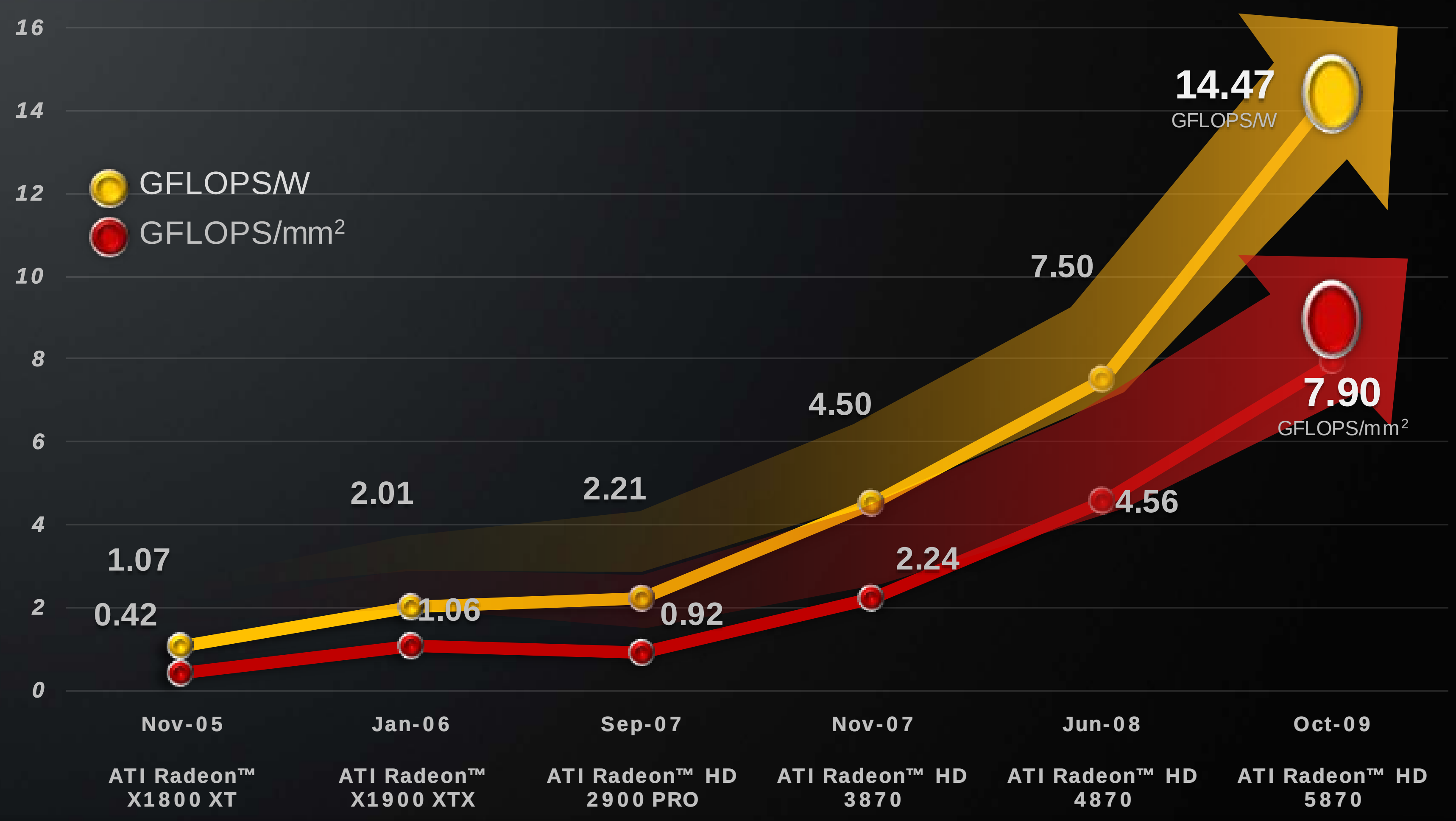


GPU Processing Performance Trend



AMD
The future is fusion

GPU Efficiency



AMD Accelerated Parallel Processing (APP) Technology is…

AMD Accelerated Parallel Processing TECHNOLOGY

Heterogeneous: Developers leverage AMD GPUs and CPUs for optimal application performance and user experience

High performance: Massively parallel, programmable GPU architecture delivers unprecedented performance and power efficiency

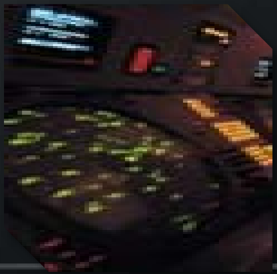
Industry Standards: OpenCL™ enables cross-platform development

Gaming

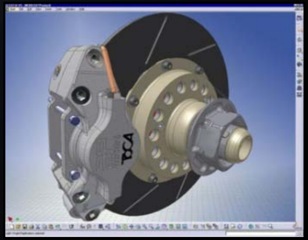
Digital Content Creation

Productivity

Sciences



Government



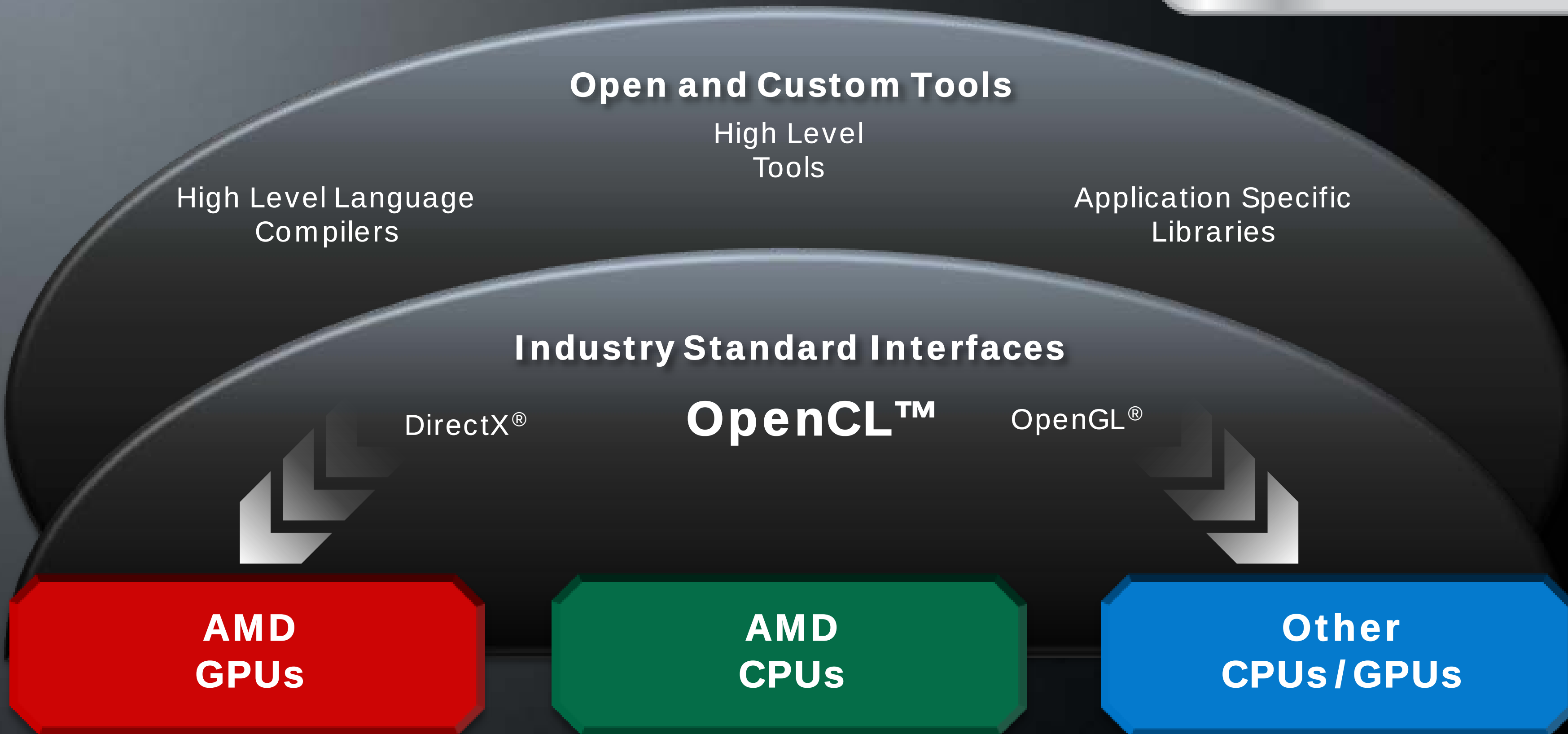
Engineering

fusion

AMD
The future is fusion

Moving Past Proprietary Solutions for Ease of Cross-Platform Programming

AMD Accelerated Parallel Processing TECHNOLOGY



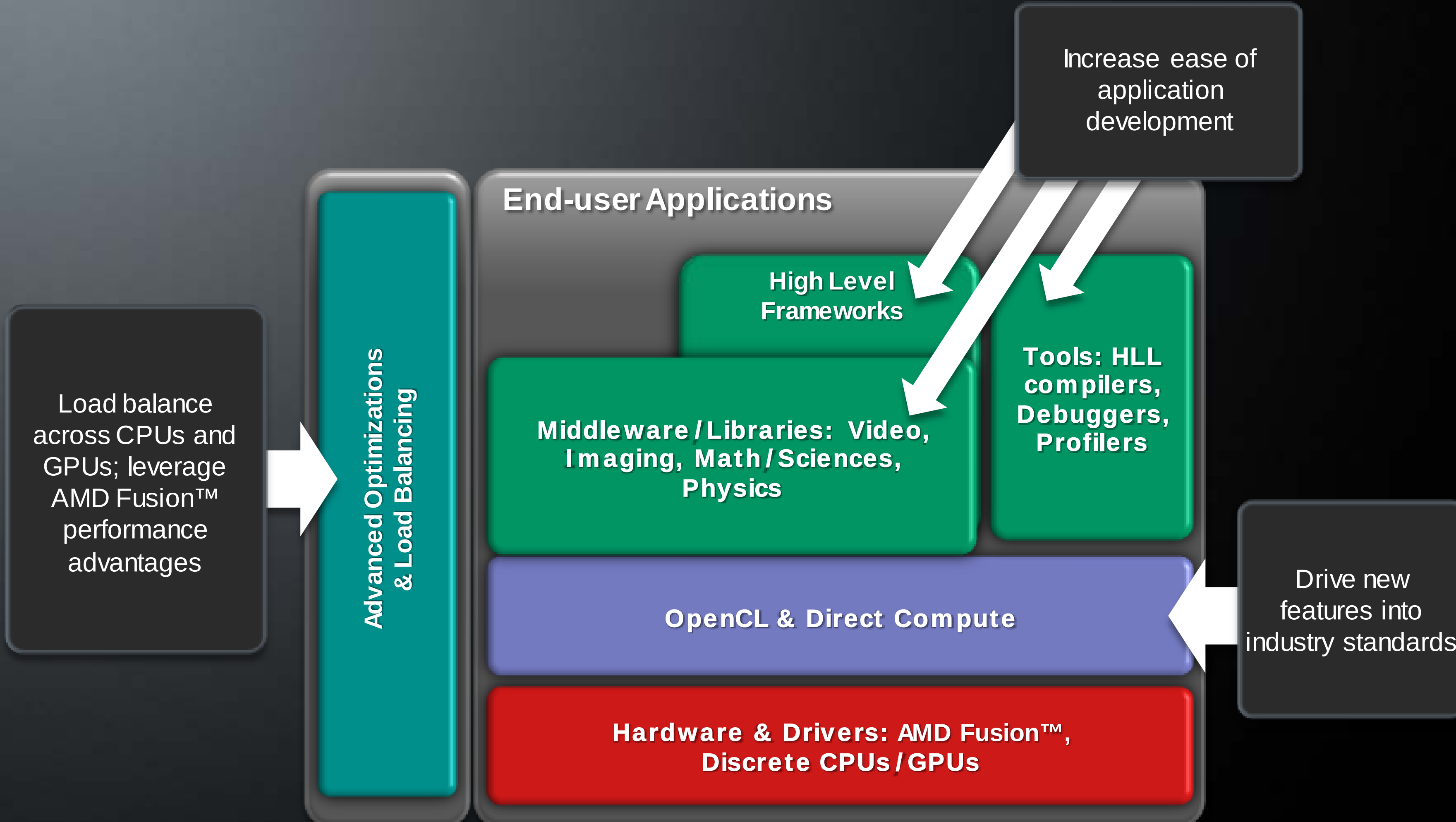
OpenCL -

- Cross-platform development
- Interoperability with OpenGL and DX
- CPU/GPU backends enable balanced platform approach

Heterogeneous Computing:

Next-Generation Software Ecosystem

AMD Accelerated Parallel Processing TECHNOLOGY



AMD The future is fusion

AMD Balanced Platform Advantage

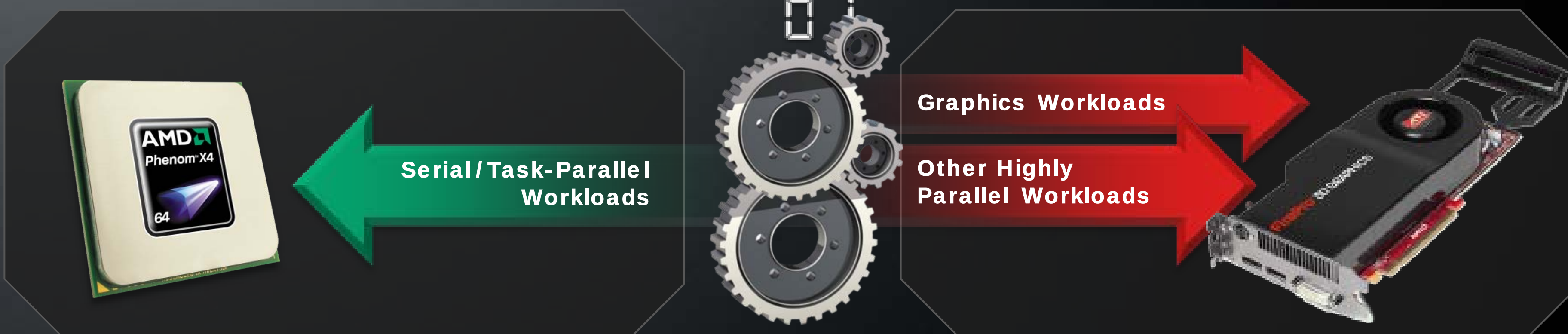
AMD Accelerated Parallel Processing TECHNOLOGY

CPU is excellent for running some algorithms

- Ideal place to process if GPU is fully loaded
- Great use for additional CPU cores

GPU is ideal for data parallel algorithms like image processing, CAE, etc

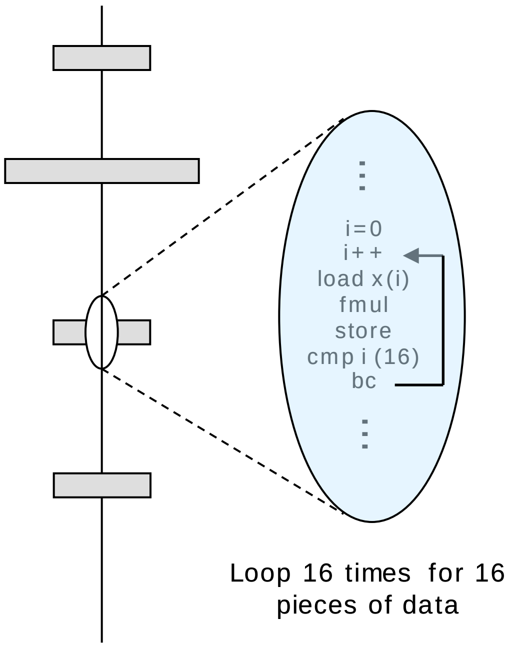
- Great use for AMD Accelerated Parallel Processing (APP) technology
- Great use for additional GPUs



Delivers advanced performance for a wide range of platform configurations

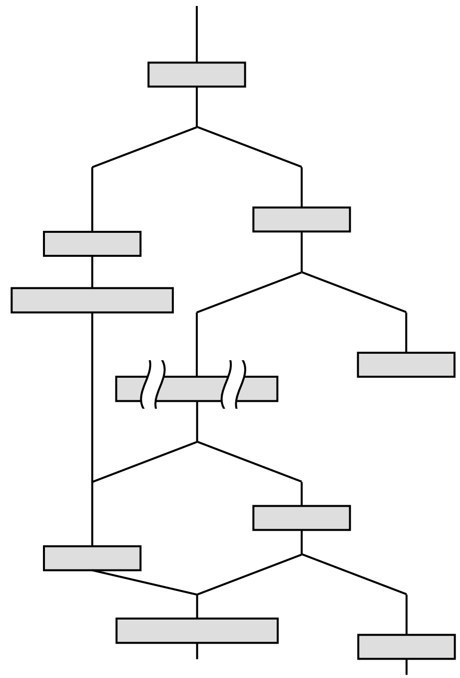
Challenges: *Extracting Parallelism*

Fine-grain data parallel Code



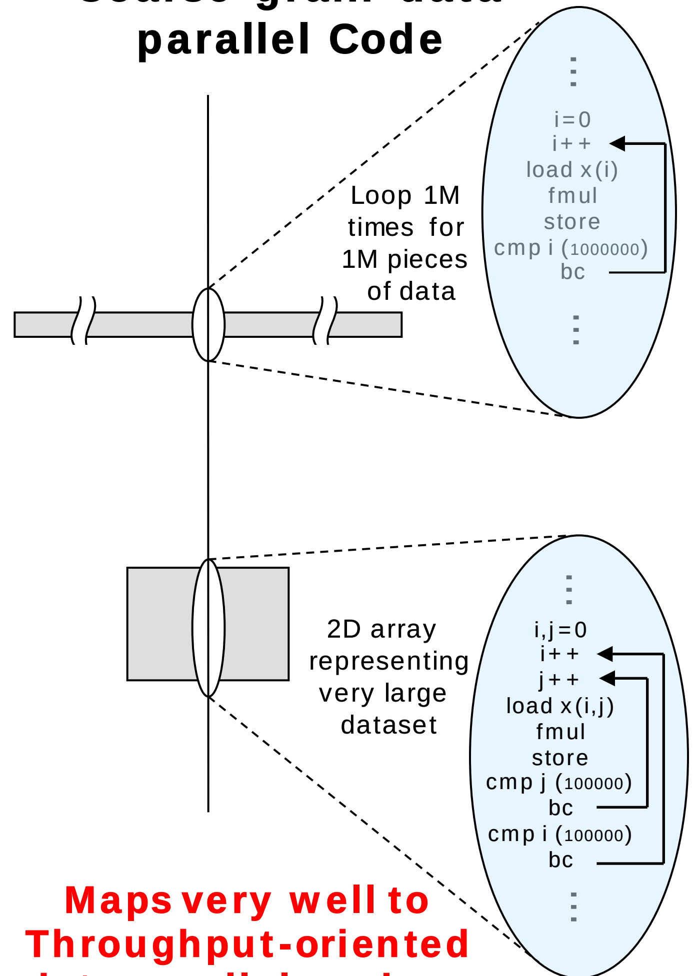
Maps very well to integrated SIMD dataflow (ie: SSE)

Nested data parallel Code



Lots of conditional data parallelism. Benefits from closer coupling between CPU & GPU

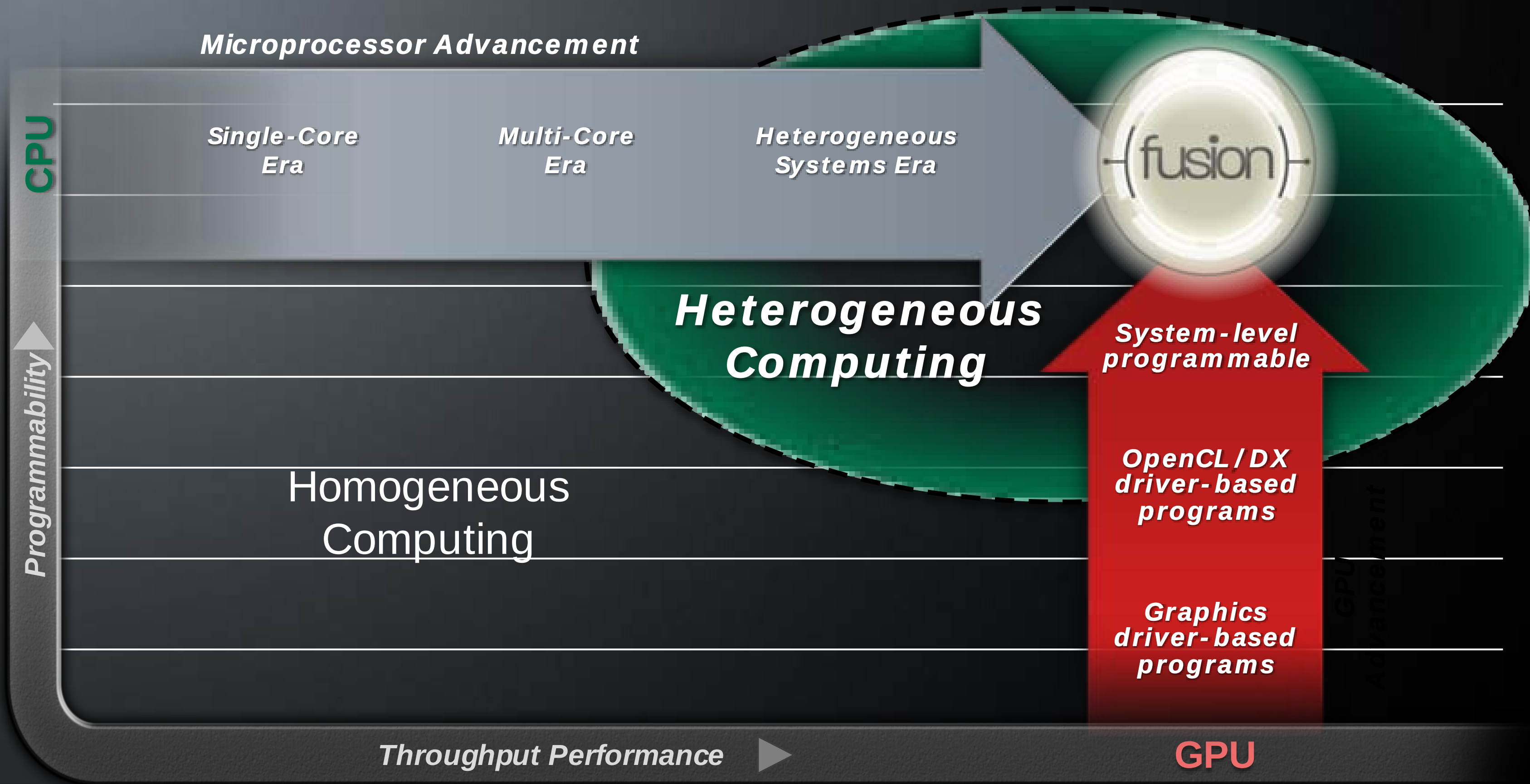
Coarse-grain data parallel Code



Maps very well to Throughput-oriented data parallel engines

A New Era of Processor Performance

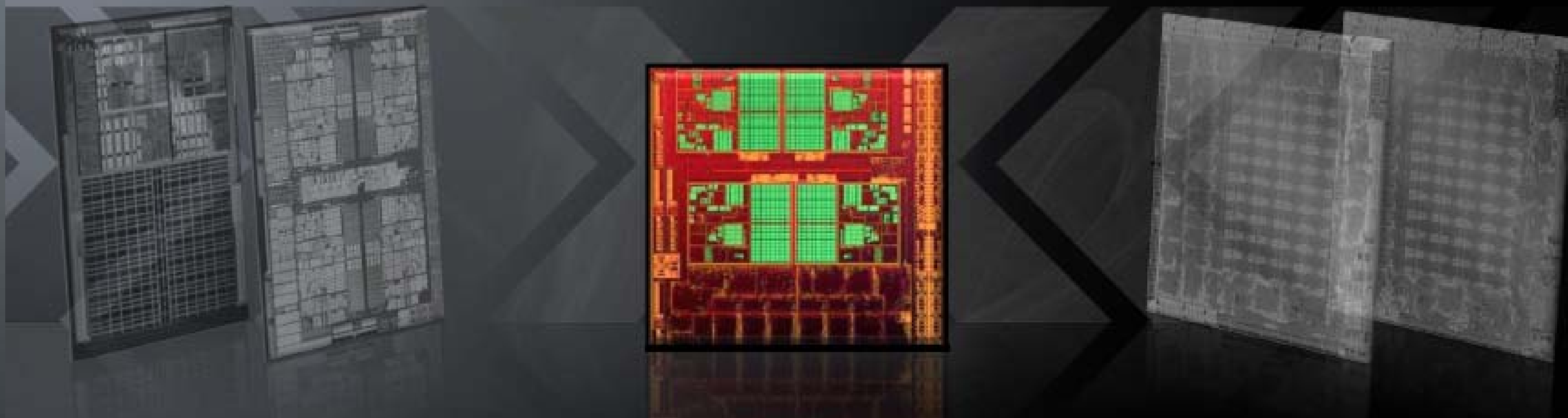
AMD Accelerated Parallel Processing TECHNOLOGY



Now the AMD Fusion Era of Computing Begins

AMD Accelerated Parallel Processing
TECHNOLOGY

Today



- ~1 billion transistors @32nm in one design
- APU: Fusion of CPU & GPU compute power within one processor
- Significantly enhances active/ resting battery life
- High-bandwidth I/O

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

This presentation contains forward-looking statements concerning AMD and technology partner product offerings which are made pursuant to the safe harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," “strategy,” “roadmaps,” "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations.

ATTRIBUTION