

2015 IEEE GlobalSIP Panel Discussion

Algorithms vs. Architectures: Opportunities and Challenges in Multicore/GPU DSP

Panelists:

Lee Barford, Keysight Technologies, US

Paul Blinzer, AMD, US

Joe Cavallaro, Rice University, US

Hong Jiang, Intel, US

Nick Moore, MathWorks, US

Yinglong Xia, IBM TJ Watson Research Center, US

Moderator:

Gwo Giun (Chris) Lee, National Cheng Kung Univ., TW

Traveling Abroad !!!

A professor traveled abroad for a meeting. Impressed by the sight of the beautiful night view, he decided to send a text message to his rather TOUGH wife:

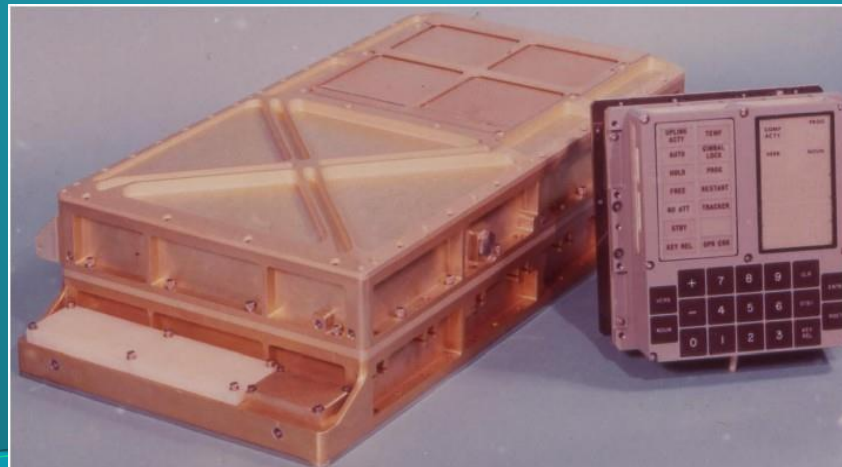
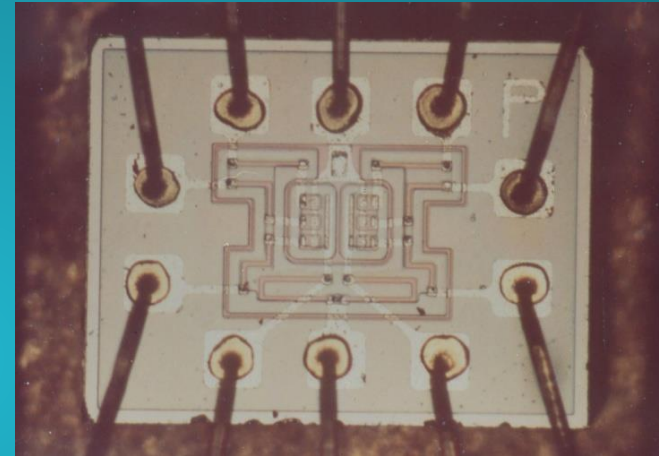
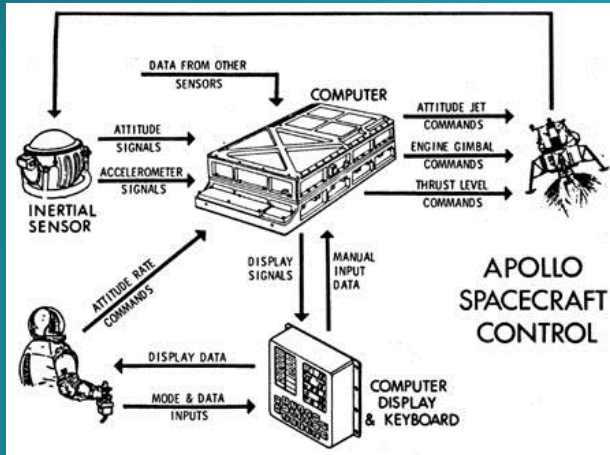
I've had a really wonderful night, and I wish so much you were here

However, the last letter "e" was accidentally omitted... and we lost contact of him after he returns home. No one knows of his whereabouts even till now!

The Math Bridge in Cambridge University



Apollo Navigation Computer



Data Gets Ever Larger

Marshall McLuhan (1960's):

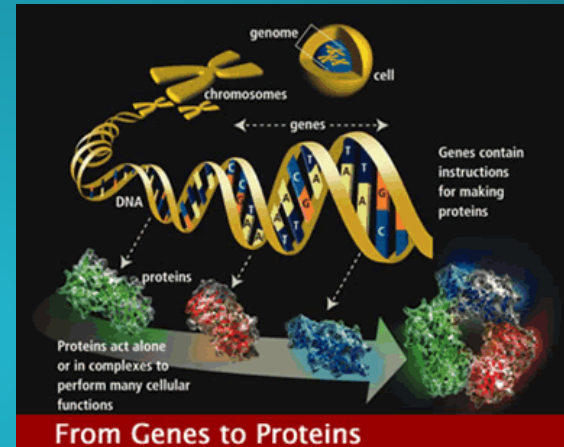
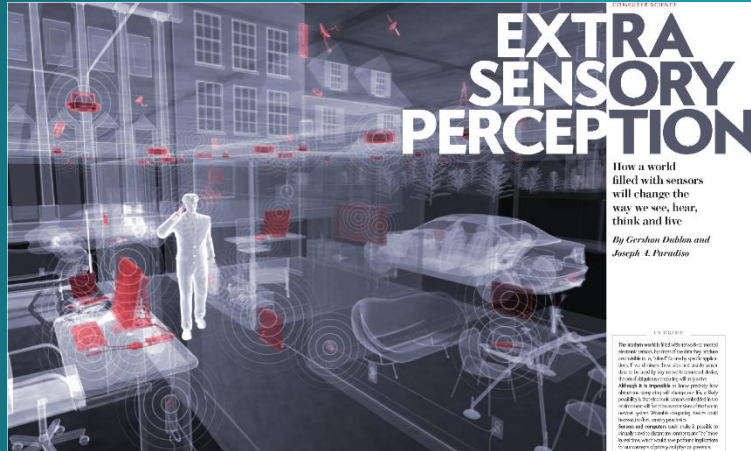
Electronic Media, primarily Television being extension of human nervous systems



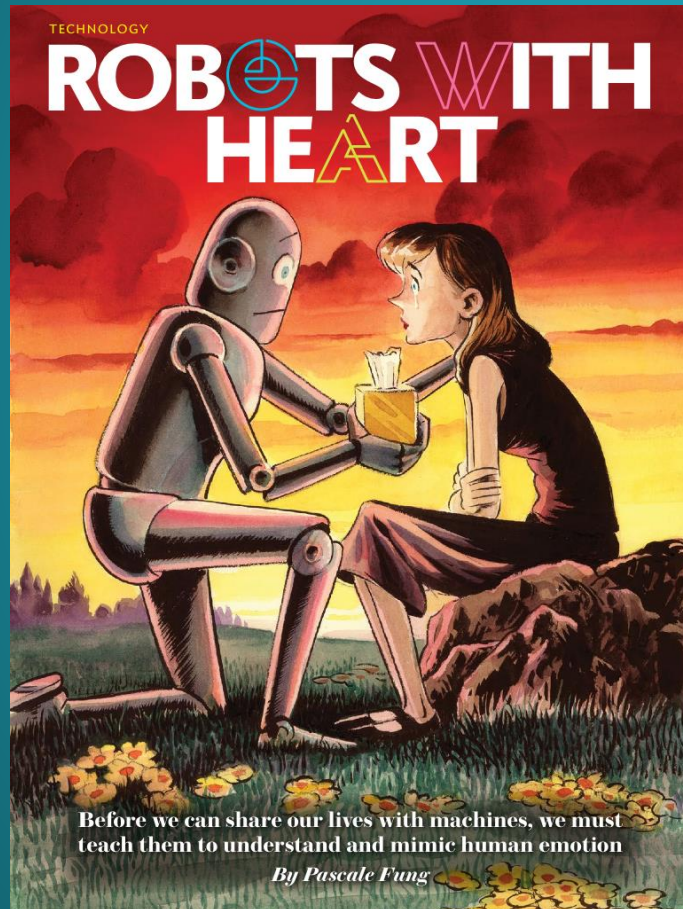
Today:

Extending outwards even further with multiple sensors interconnected. When going deeper inwards into the human body with huge data from human brain and human genome

Reaching Out Even Further via IoT and Going in Ever Deeper to the Human Brain and Genome

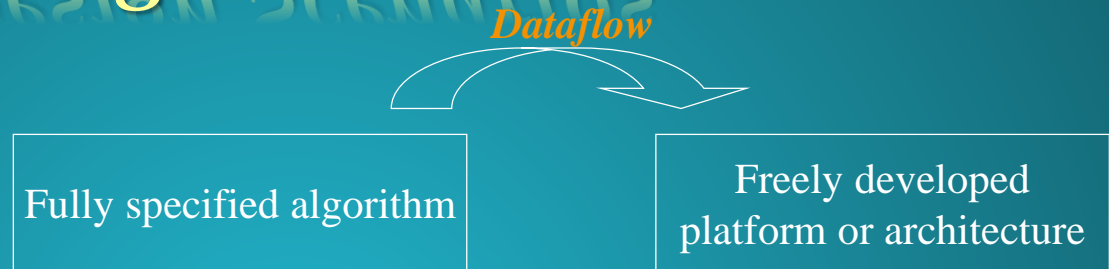


Algorithms Get Ever More Complicated Towards Automation

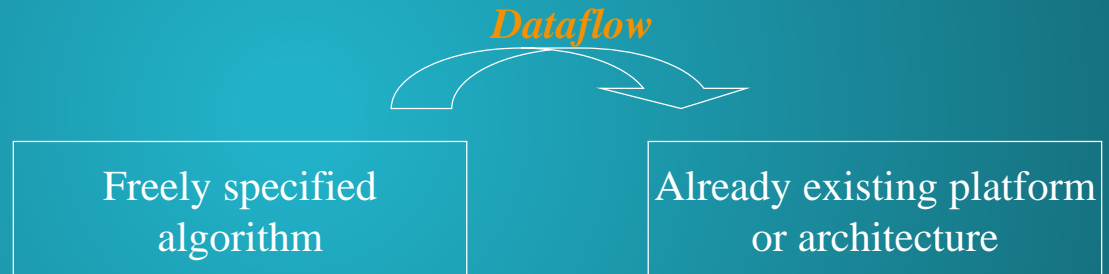


Generic Signal Processing System Design Scenarios

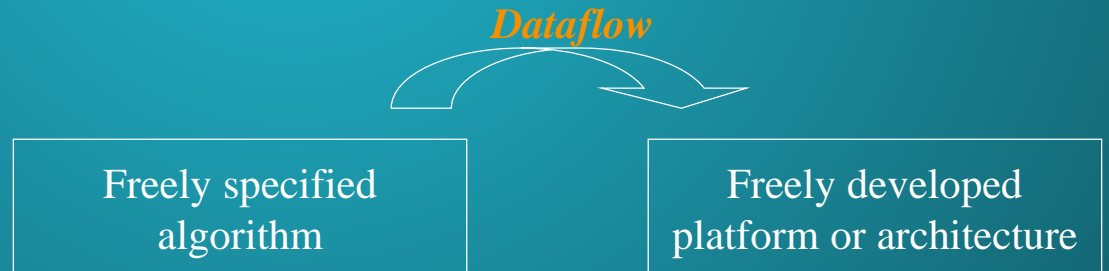
Scenario I:



Scenario II:



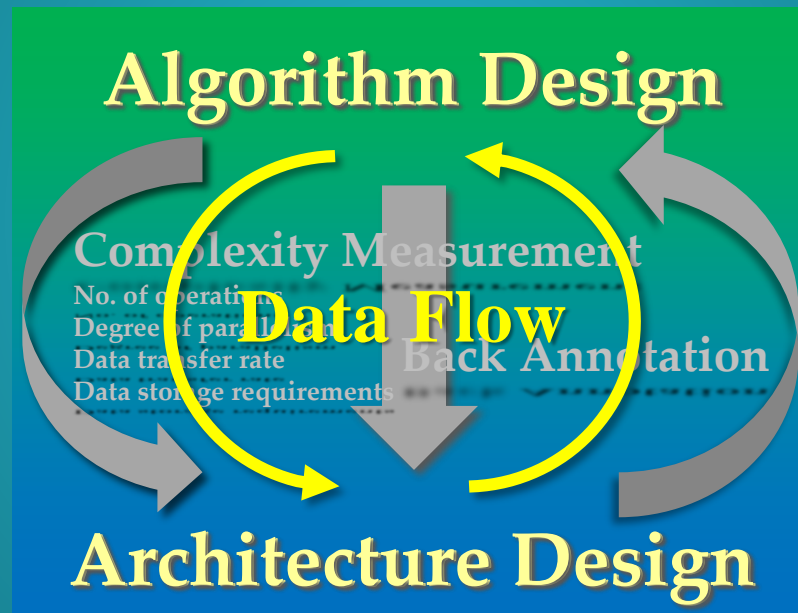
Scenario III:



subject to usual project constraints in terms of **performance per unit silicon area**, **flexibility**, **power consumption**, etc.

Algorithm/Architecture Co-Exploration

Algorithm/Architecture Co-Exploration



G. G. (Chris) Lee, Y.-K. Chen, M. Mattavelli, and E. S. Jang, "Algorithm/Architecture Co-Exploration of Visual Computing: Overview and Future Perspectives," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 19, Iss. 11, pp. 1576-1587, Nov. 2009.

New Design Paradigm: Moving from programming to design and beyond... Big Data

Wirth from ETHZ (1975):

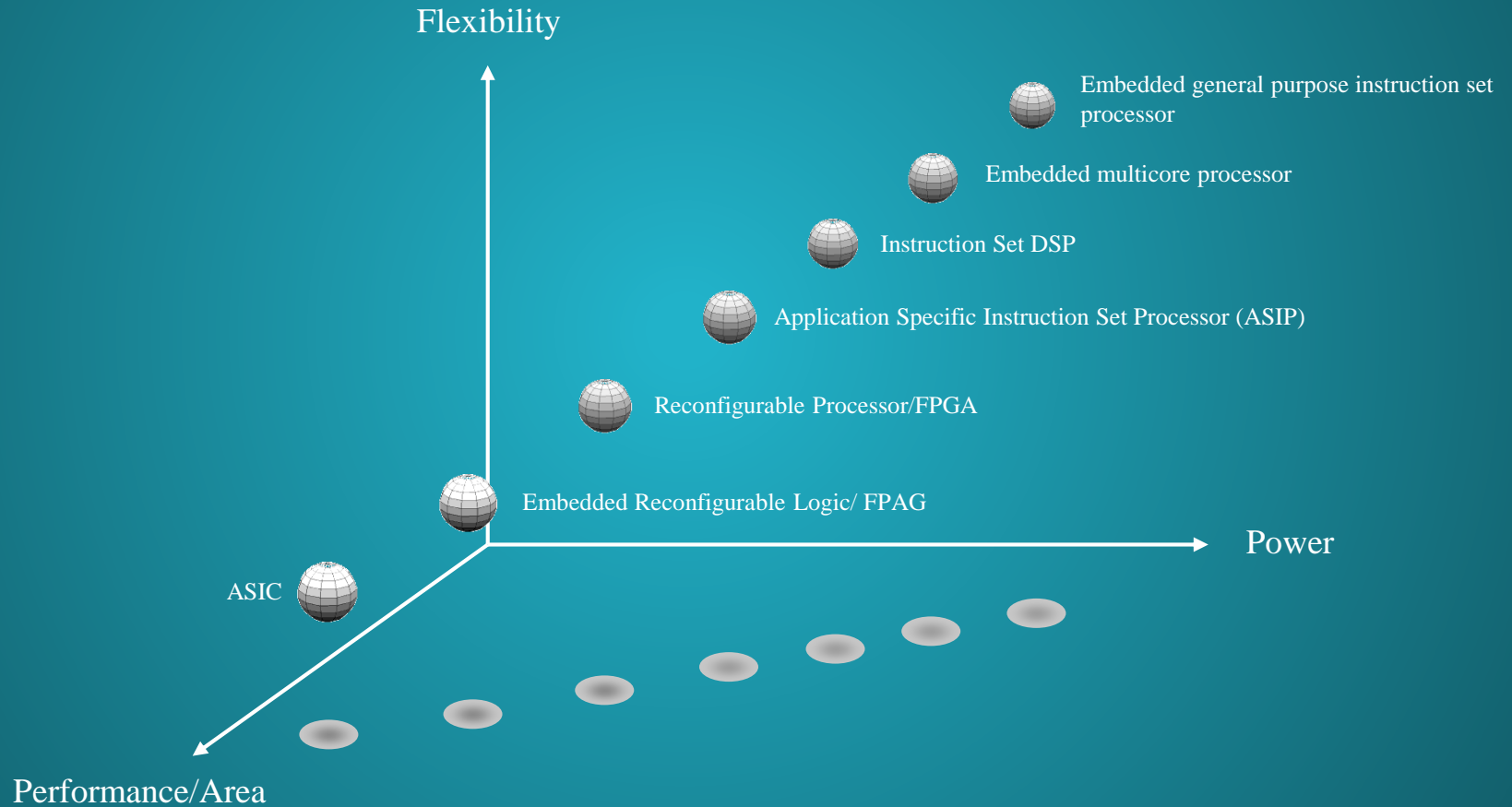
Programming = Algorithm + Data Structure



Lee from NCKU (2007):

Design = Algorithm + Architecture

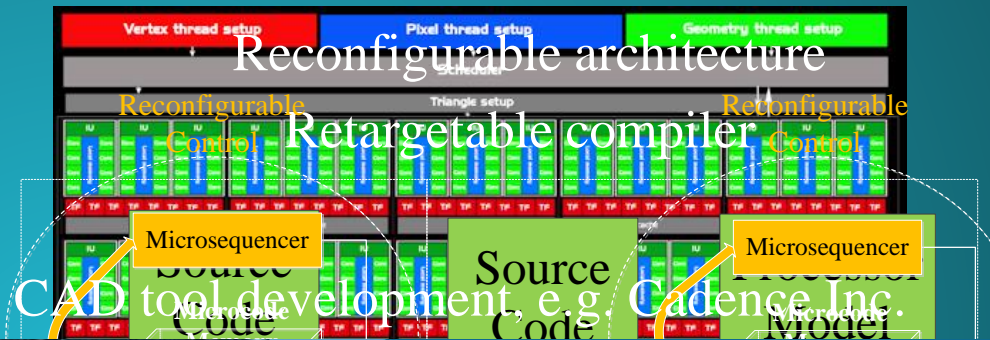
Architectural Platforms Before Cloud



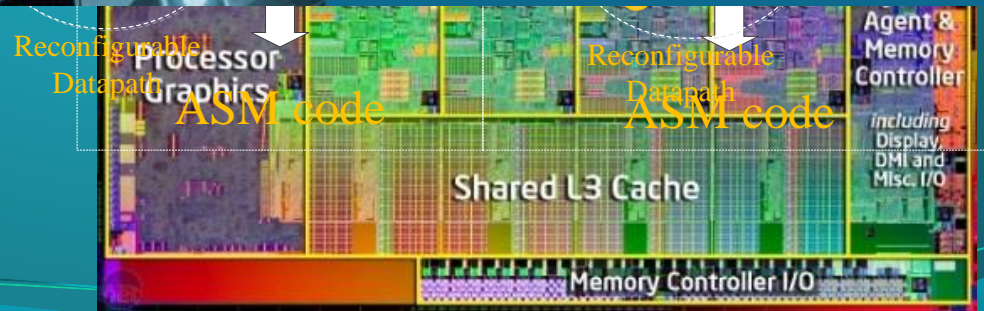
Intelligent Parallel/Reconfigurable Computing

Homogeneous multicore processor (Nvidia GT200)

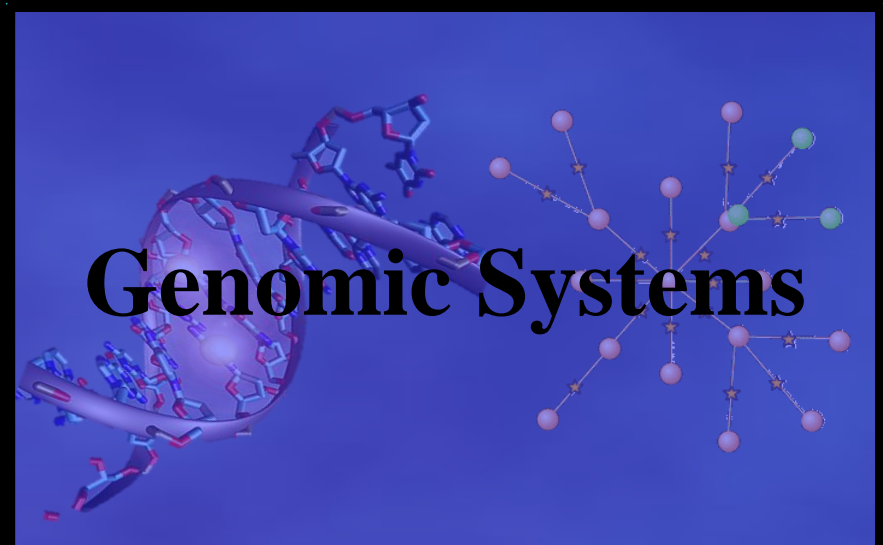
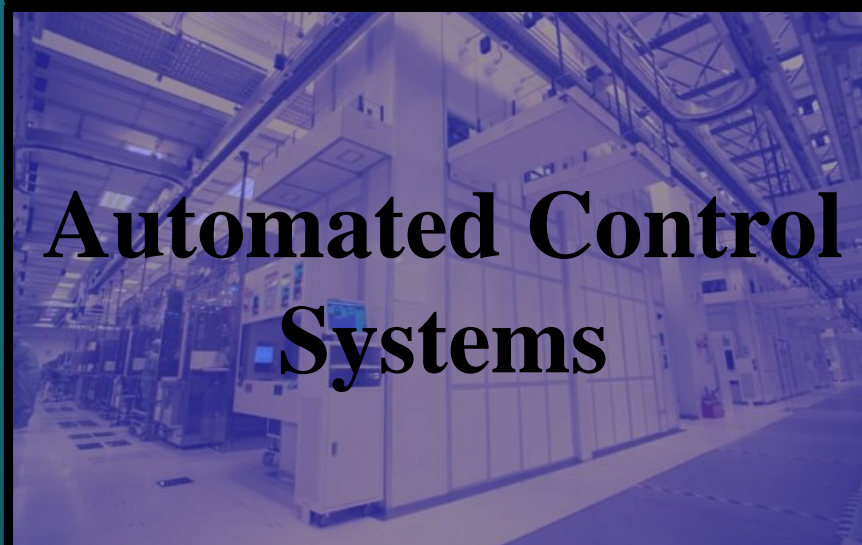
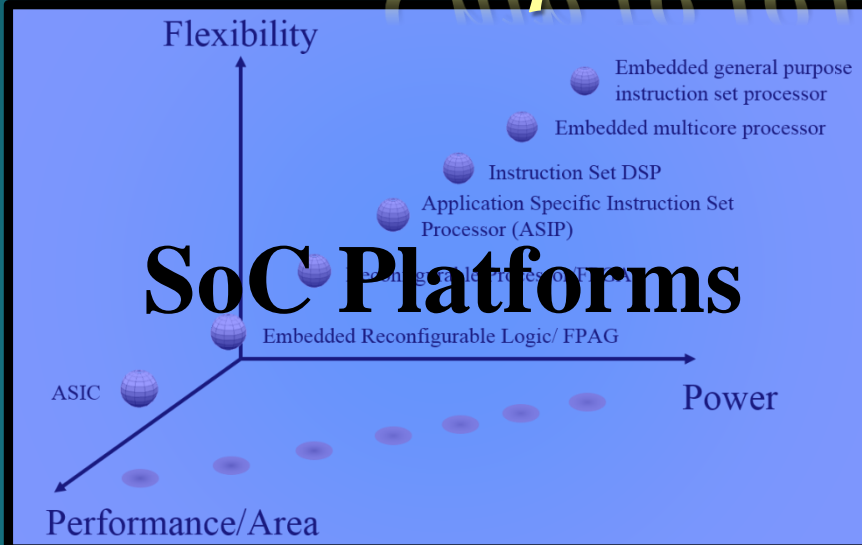
- Parallelism quantification
 - Intelligent tasks/processes and resource allocation
- Data transfer and data storage analysis
 - Intelligent communication protocol controlling
 - Intelligent storage management



Accelerating system-level design, integration, and verification

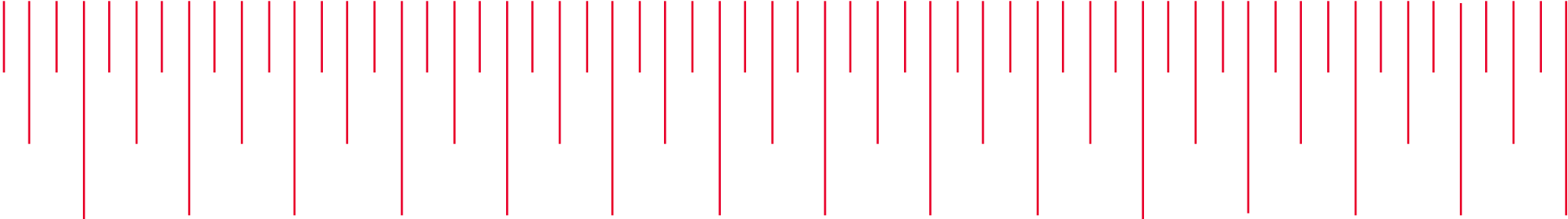


System Platforms: From System-on-Chip to IoT and Cloud



Computer, Communication, Control & Care

Thanks for Your Attention!



Algorithms vs. Architectures: Opportunities and Challenges in Multicore/GPU DSP: Introduction for Panel

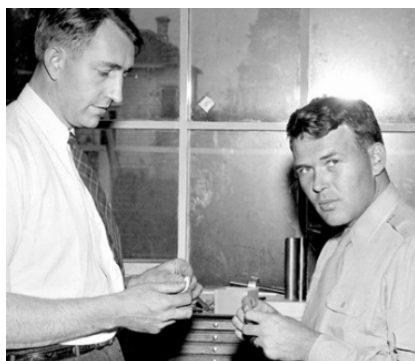
Updated December 17, 2015

Lee Barford
Fellow
Keysight Laboratories

A Brief History of Keysight

We believe in “Firsts”

Bill Hewlett and Dave Packard’s vision launched Silicon Valley and shaped our passion for “firsts” 75 years ago. Today we are committed to provide a new generation of “firsts” – software-oriented solutions – that create value for our investors and valued insights for our customers



1939–1998: Hewlett-Packard years

- A company founded on electronic measurement innovation
- Grew successfully as a Premier Test and Measurement company.
- HP Introduced the early computers and printers, and captured huge growth



1999–2013: Agilent Technologies years

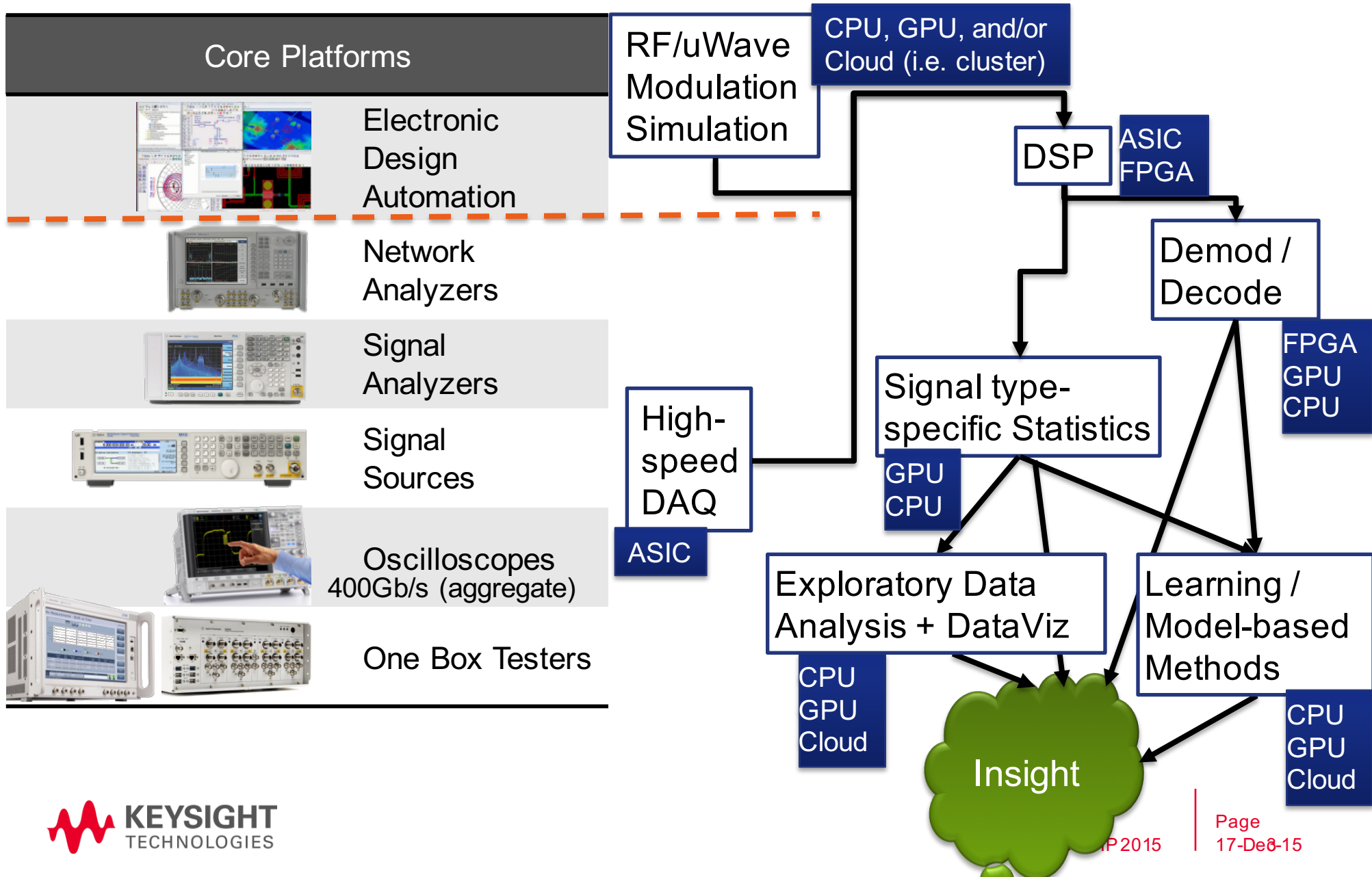
Agilent Technologies spun off from HP, became the World’s Premier Measurement Company. In September 2013, Agilent announced the spinoff of its electronic measurement business.



2014: Keysight begins operations

Keysight Technologies was spun off from Agilent on 1st November 2014, as an independent company focused 100% on the electronic measurement industry.

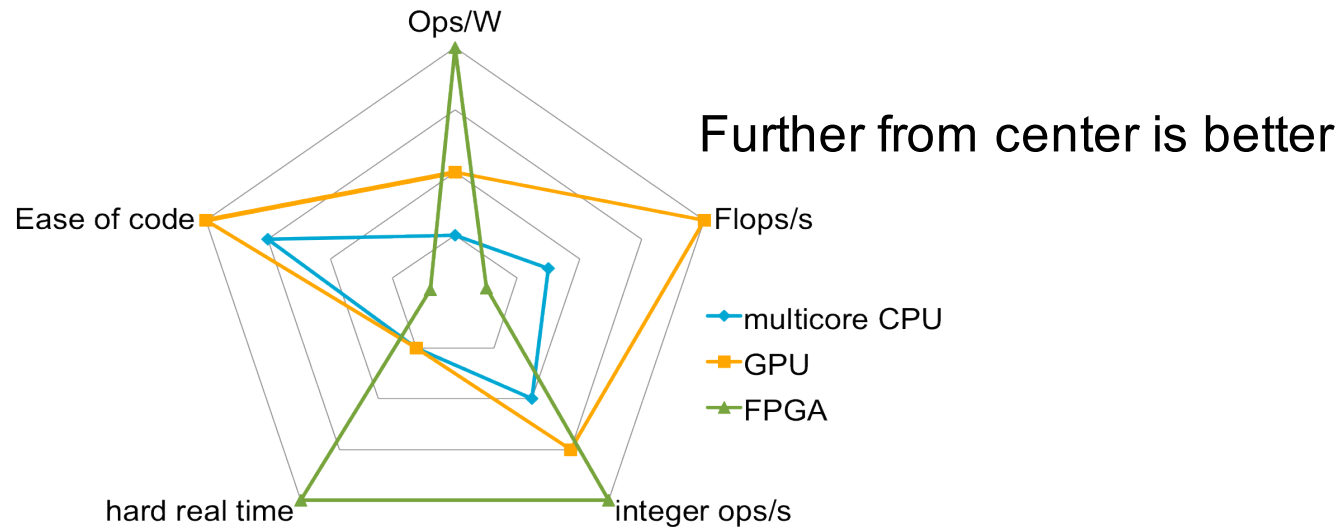
High-Performance Measurement Previews the Future of Big-Data Fused Signal Processing



Hybrid processing is key

Extrapolating to the vision of learning from a plethora of signals

- Needed to achieve economically desirable price/performance/power/R&D cost



- Slowing of Moore's Law → Maybe need proliferation of architectures on a single CPU to maintaining growth of computing speed at acceptable wattage
 - cf. Andrew Chien's 10X program @ U. of Chicago
- Increased use of hybrid processing increases software and integration challenges
 - Multiple code bases for same functionality, efficiency of data transfer, walls between communities of expertise within the same enterprise

Veracity in the extremely signal-rich environment

Epistemic reliability requires analogs of “3 pillars of metrology*”

- **Uncertainty:** Model for expressing posterior $P(\text{physical quantity } Q \mid \text{data } d)$ appropriate to the physical quantities, d , and the way d is obtained
- **Calibration:** A two-step process
 1. Obtain data d_1 from known Q_1 , d_2 from known Q_2 , ..., d_N from known Q_N
 - Q_i 's called calibration standards
 2. Define algorithm giving $P(Q \mid d, Q_1, \dots, Q_N, d_1, \dots, d_N)$ for unknown Q
- **Traceability:** Calibration uses calibration standards that are calibrated to other standards that are calibrated to other standards ... that are calibrated to experiments than define the SI units (or the Grand kg in Paris)
- The 3 Pillars permit valid inferences and reliable learning from the posteriors
- Provided for in instruments and periodically renewed in instrument service depots
- How to do this to the proper level of formality, flexibly yet reliably, in the envisioned signal rich world with dynamically-applied data-mining and learning?

BACKUP SLIDES

Keysight at a Glance

REVENUE IN FY14	\$2.9 billion
EMPLOYEES	9,600
PRESIDENT and CEO	Ron Nersesian
GLOBAL HEADQUARTERS	Santa Rosa, California
CUSTOMER LOCATIONS	100+ countries
MANUFACTURING AND R&D LOCATIONS	U.S., Europe, Asia Pacific
NYSE	KEYS



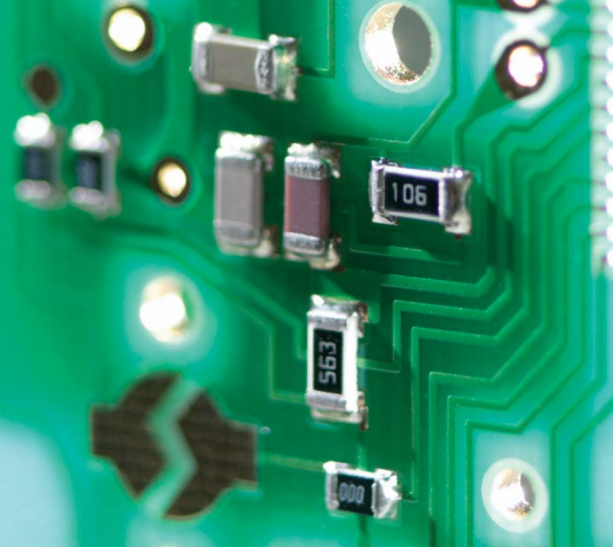
Ron Nersesian
President and CEO

Keysight in Electronic Measurement

The industry leader



Communications



Industrial, computer,
semiconductor



Aerospace/defense

FY14 \$2.9 billion revenue | 19.1% operating margin | 31% ROIC | best in class financial profile

(1) Non-GAAP measure. See reconciliation to GAAP financial measures.

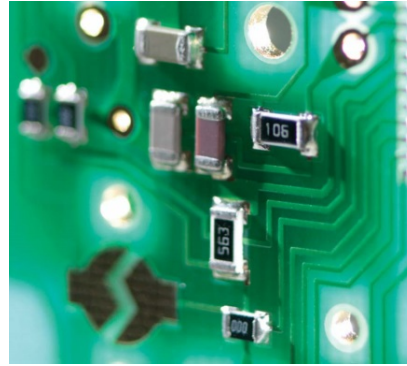
We Help Companies Unlock Insights to Succeed



Communications:

From the speed of innovation to the cost of test, time-to market pressures have never been greater.

We help companies win in the first to market race.



Industrial, Computers, Semiconductors:

Electronic content is everywhere.

Explosive growth calls for a proven partner. We help customers across design, verification and manufacturing to installation and maintenance.



Aerospace Defense:

Where there's no room for compromise, we help customers reduce risk.

We help customers update their radar, satellite and communication systems.

Keysight Customer Support and Service

Founded on deep customer relationships and trust

- **Global Reach & Capability**

- Service centers in 30 countries repair and calibrate customer test equipment
- Consistent support for multi-national customers at 50+ sites worldwide

- **Broad Service Offering**

- Designed and delivered by a global team of experts
- Mobile on-site calibration services
- Trade-in and certified used product sales

- **Deep Domain Expertise**

- Experts in the science of electrical and physical dimensional measurements
- Affiliated with 35 calibration standards bodies in 17 countries



Cutting down the noise Processing data for relevance

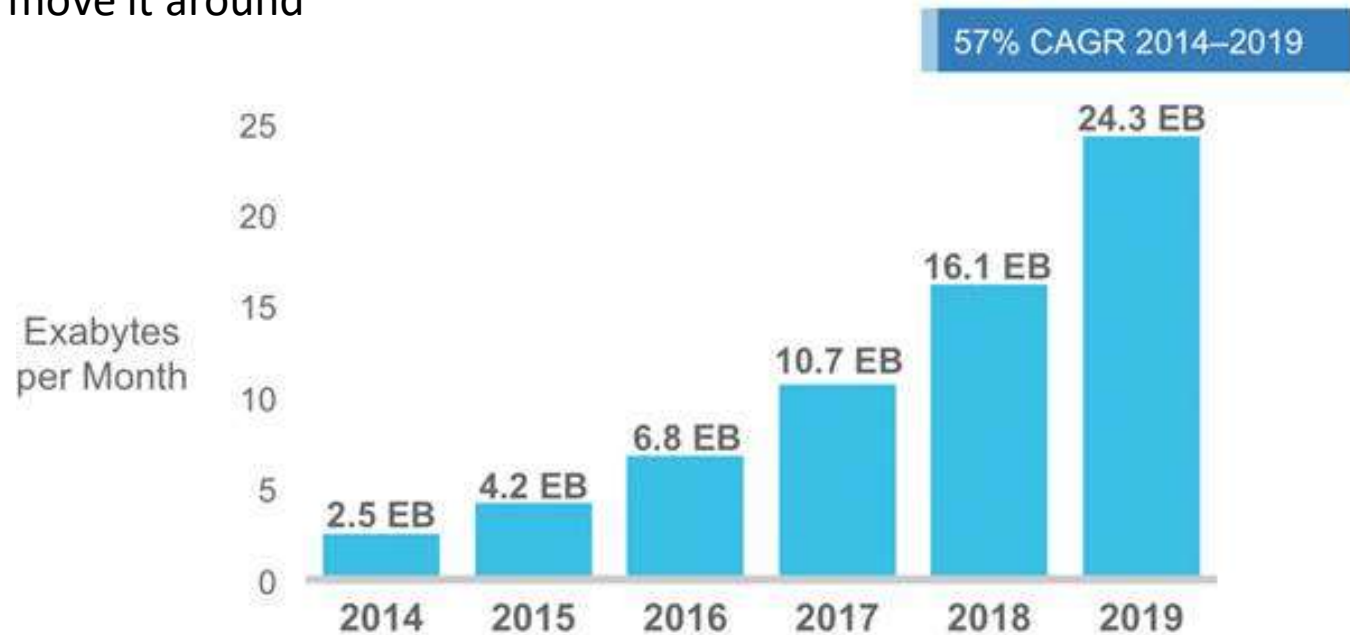
Paul Blinzer, Fellow System SW, AMD Inc
System Architecture Workgroup Chair, HSA Foundation

The amount of data created grows and grows...

- data is created, managed, stored and retrieved at ever increasing amounts
 - News, video, audio, environmental sensors for spatial or visual awareness, ...
- The challenge becomes to identify & process the relevant information
 - break it down into pieces that a human and “regular software” can process
 - Process it close to the origination point, don't move it around

The human sensory capacity to process information is both extensive and limited

- Human senses have highly sophisticated pattern matching capabilities to process the environment
- Semantic, higher-level “data compression” needed and worked on that models human perception
- Algorithmic processing on dedicated accelerators needed



source: Cisco forecast

How to manage all?

- Search engines
 - First line of defense in the internet data age
 - Each of the major search engine providers use an enormous amount of computing power to retrieve potentially relevant information
 - Using a lot of personalized & contextual information
 - Still a lot of noise in the search results – if it's not on the first page, it's lost
- Social Media news feeds
 - The human element in data retrieval, you get the news and info that your friends are interested in, likely interesting you
- The rise of the digital assistant: Siri, Cortana, OK Google, Alexa
 - Deep personalized user context knowledge attempts to predict relevant info
 - Only few choices are presented, but it has to be relevant info
- Fundamentally based on Deep Neural Network algorithms
 - Training for relevant data, extrapolation and feedback
 - A lot of the “analogue” information is broken down into relevant data
 - Benefits from offloading to dedicated, highly parallel processors

The perception of reality

- Visual data presentation (e.g. through VR) one of the major ways to present huge amounts of data, but needs other, sensory input to align
- High selection focus -> Importance that the data retrieved is complete, accurate and relevant
- Correlation vs causation is a problem
 - Seeing patterns where there are none
 - Not only a “human problem” 😊
- Importance of data processing fail-safes
 - System design needs to take different overlapping processing algorithms into account to build redundancy
 - Different accelerator architectures need to be integrated efficiently in software



Algorithms vs. Architectures: Opportunities and Challenges in Multicore / GPU DSP

Joseph R. Cavallaro

ECE Department, Rice University, Houston, TX

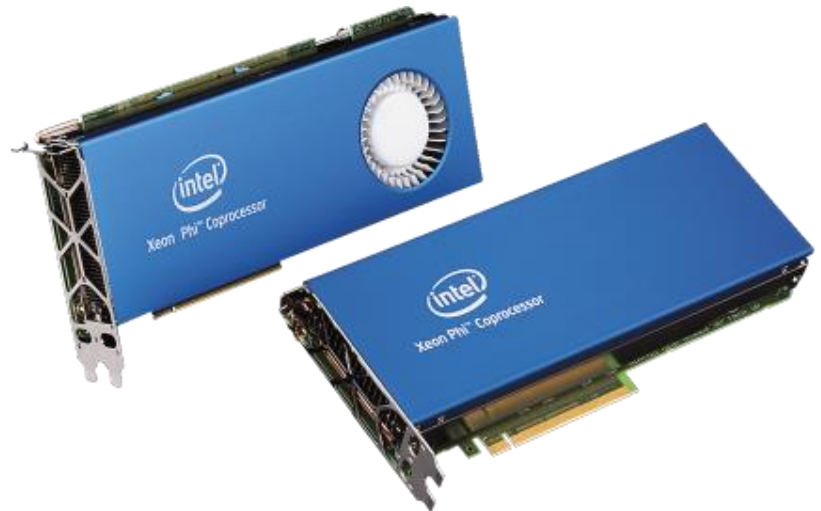
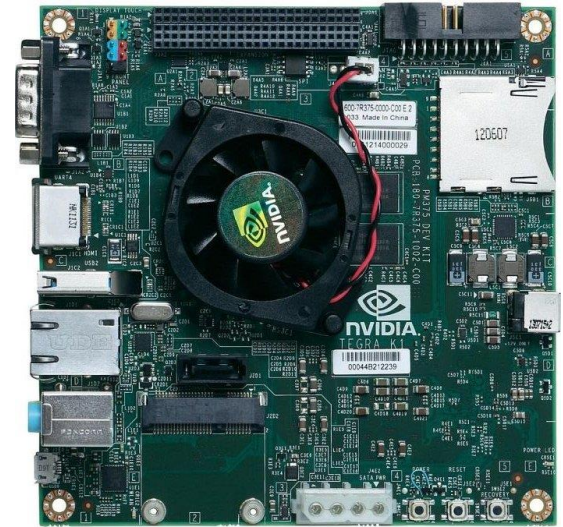
cavallar@rice.edu

Dec. 16, 2015



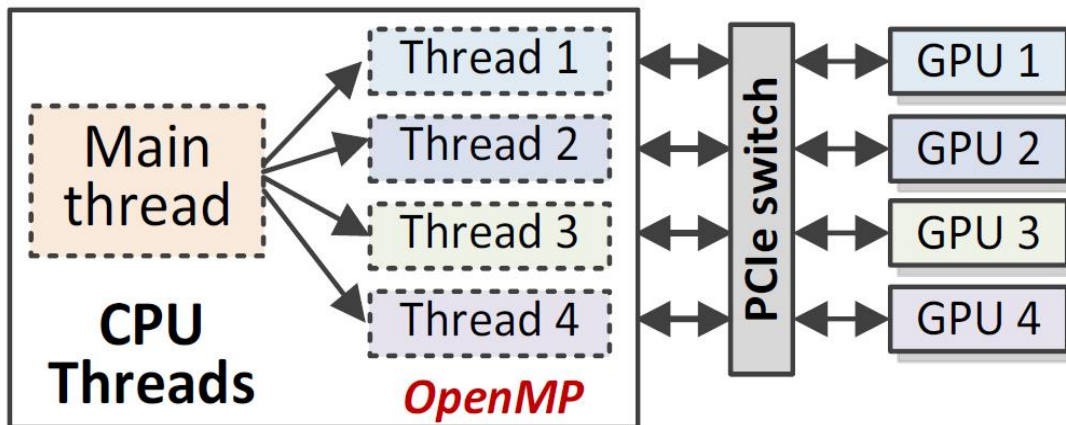
Opportunities

- Accelerators abound
- GPU – Tegra to Titan scale
- FPGA – Custom structures
- Multi-core – Intel Phi
- Tools support
- OpenCL, OpenMP ...
- CUDA, various C to HLS ...
- HSA ...



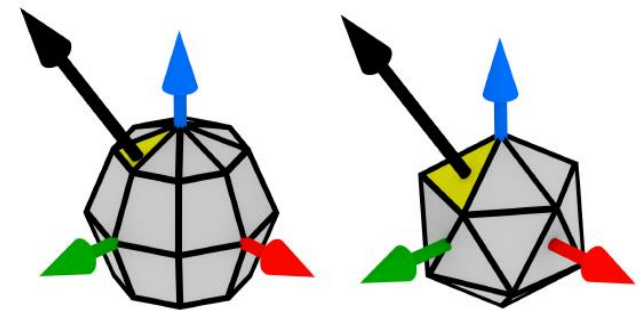
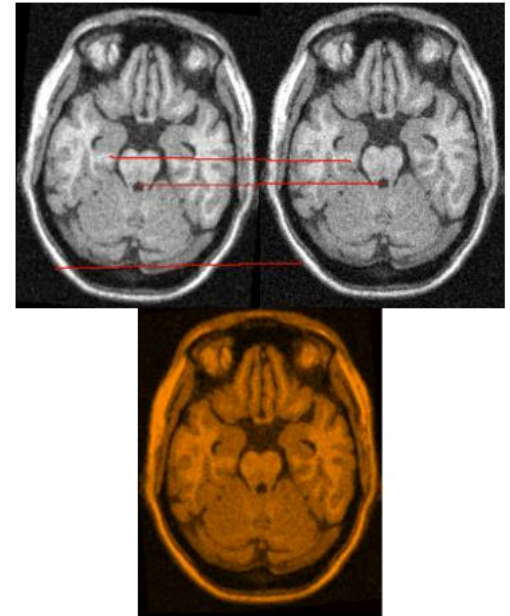
Challenges

- Hardware – Software partitioning
- Limits to algorithm parallelism - utilization
- Memory transfers from host CPU to GPU and back
- Lack of customization
- Synchronization issues



Case Study 1: MRI Imaging

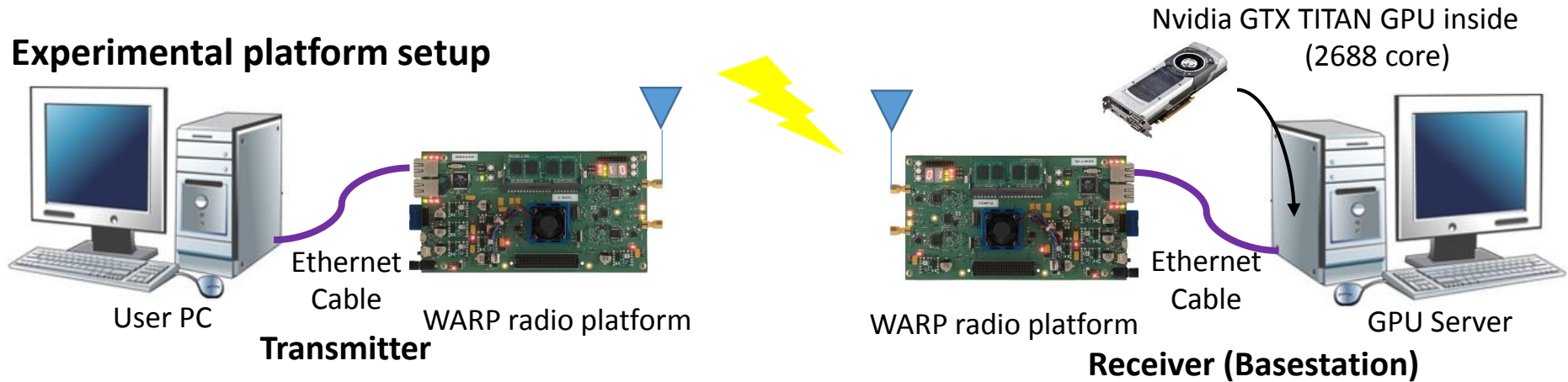
- MRI image registration challenge
- GPU acceleration of feature extraction
- 3D image sets and GPU organization
- SIFT, SURF algorithms are parallel
- Limited || at different stages
- Goal to quickly find changes over time
- Disease progression - MS



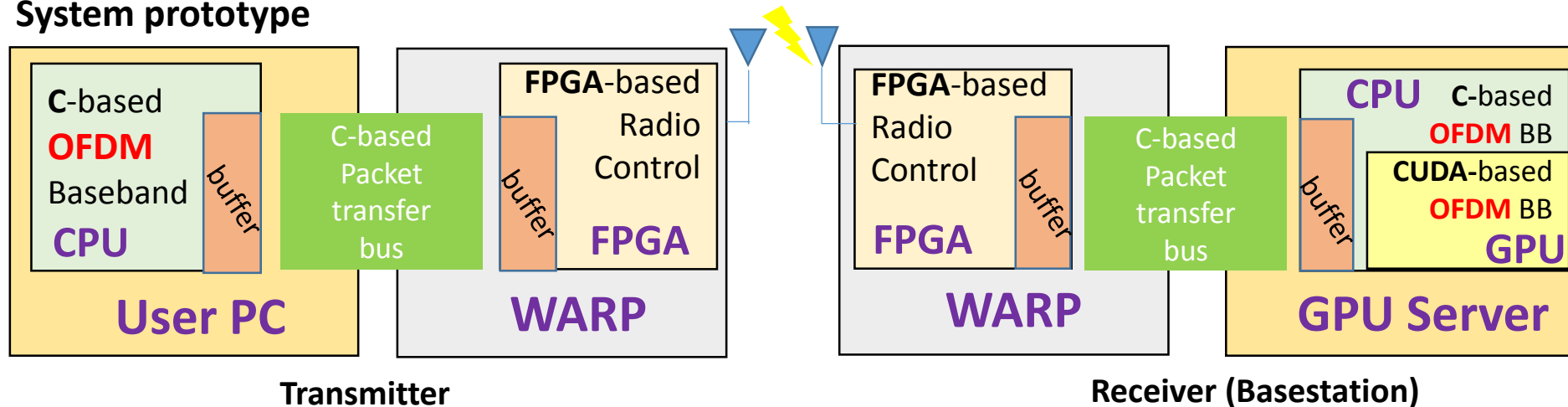
Case Study 2: a GPU-based OFDM System

- Targeting system: A SISO OFDM system for WiFi uplink
- TX: Include a user PC and another WARP to perform streaming data transmitting
Baseband processing modules are not complex and implemented in C on CPU
- RX: GPU-based software-defined basestation plays the role of receiver in this case
Some of complex modules are implement in CUDA on GPU

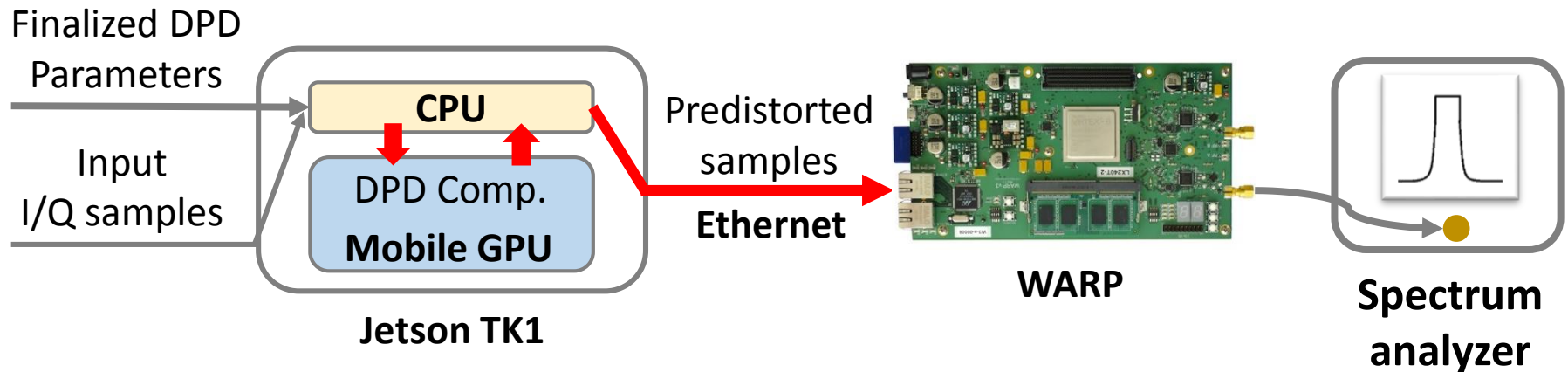
Experimental platform setup



System prototype



Case Study 3: Mobile GPU-based predistortion system



Key strategies to enhance system data-rate performance:

➤ Improve computation efficiency

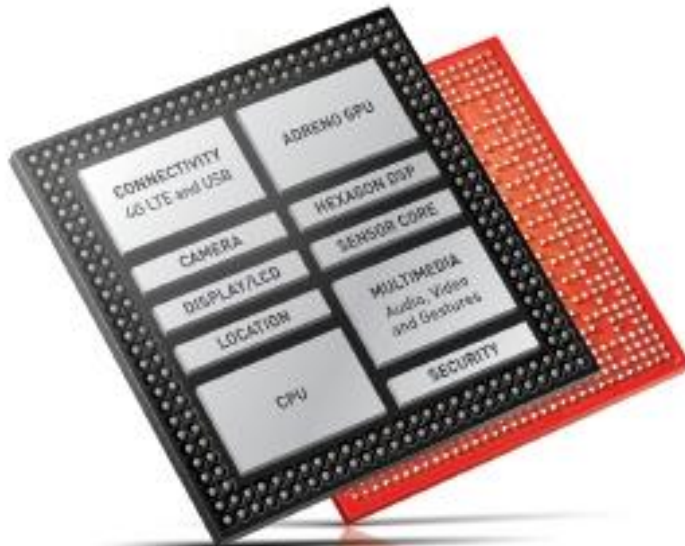
- Multi-threaded DPD computations on mobile GPU
- Memory access optimization on mobile GPU

➤ Reduce data transfer overhead

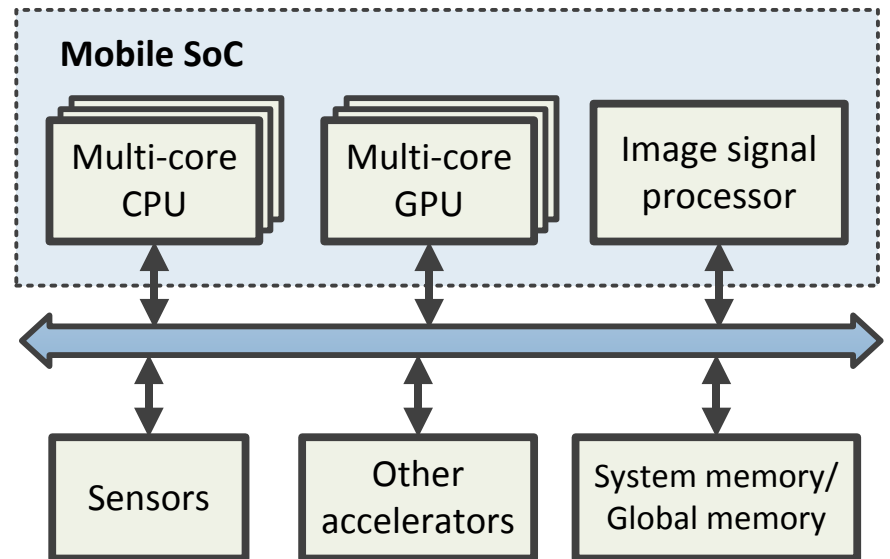
- Reduce CPU-GPU memory copy overhead in Jetson
- Reduce packet transfer overhead between Jetson and WARP

Going Forward

- Greater system integration
- Mobile SoC as possible example for larger systems
- Better design support will be critical
- HSA may be a way forward



Qualcomm Snapdragon 820





ALGORITHMS VS. ARCHITECTURES: OPPORTUNITIES AND CHALLENGES IN MULTICORE / GPU DSP

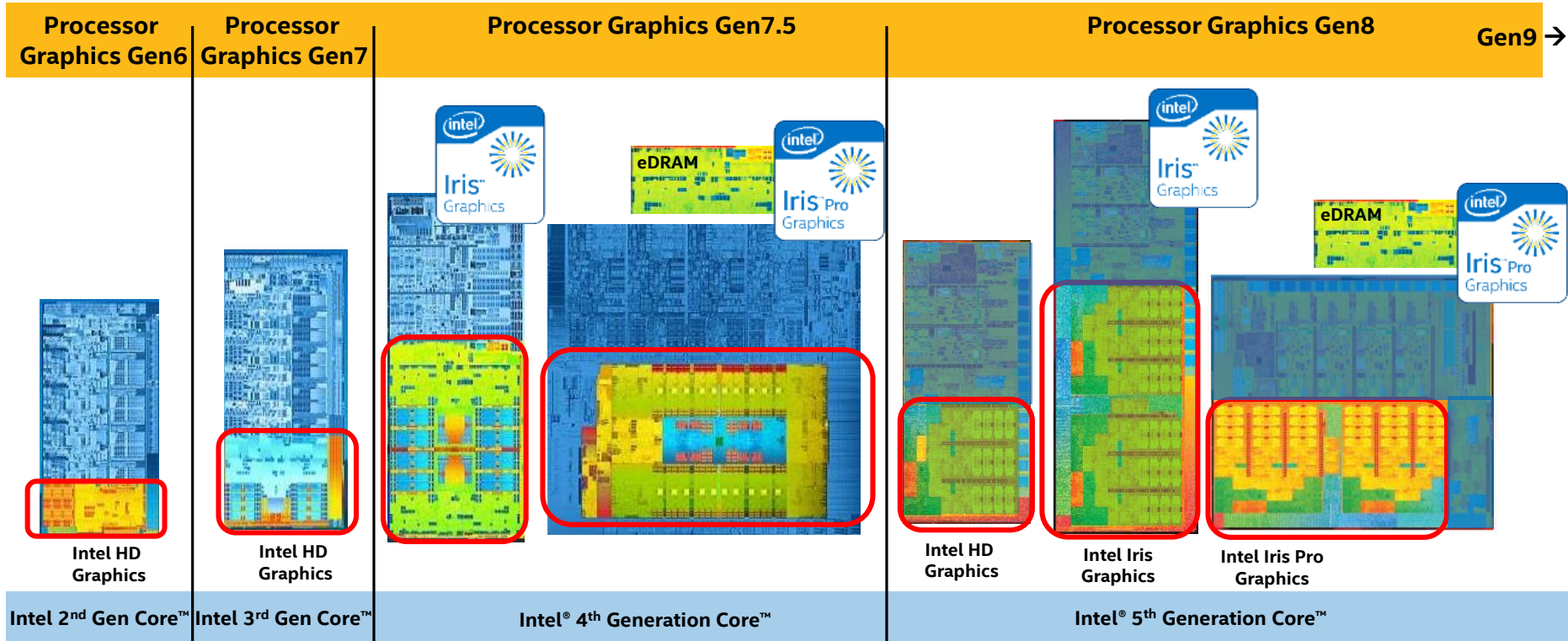
INTEL[®] MICROPROCESSORS AND PROCESSOR GRAPHICS

Dr. Hong Jiang – Intel Fellow & Director of Media Architecture

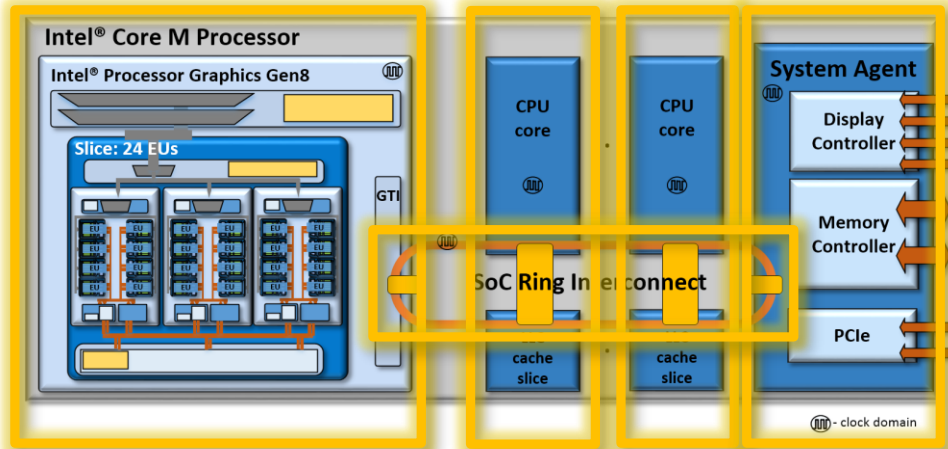
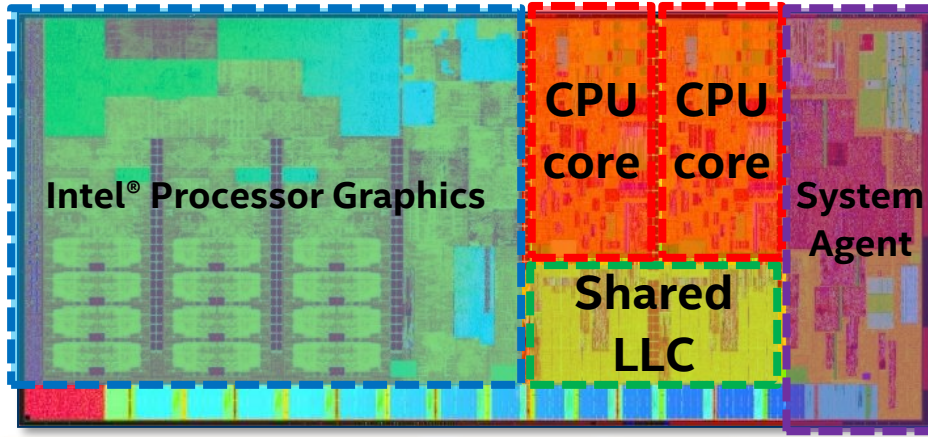
Visual & Parallel Computing Group, Intel Corporation

IEEE GlobalSIP 2015, Orlando, Florida, USA, Dec. 16, 2015

Processor Graphics is a Key Component of Intel Microprocessors



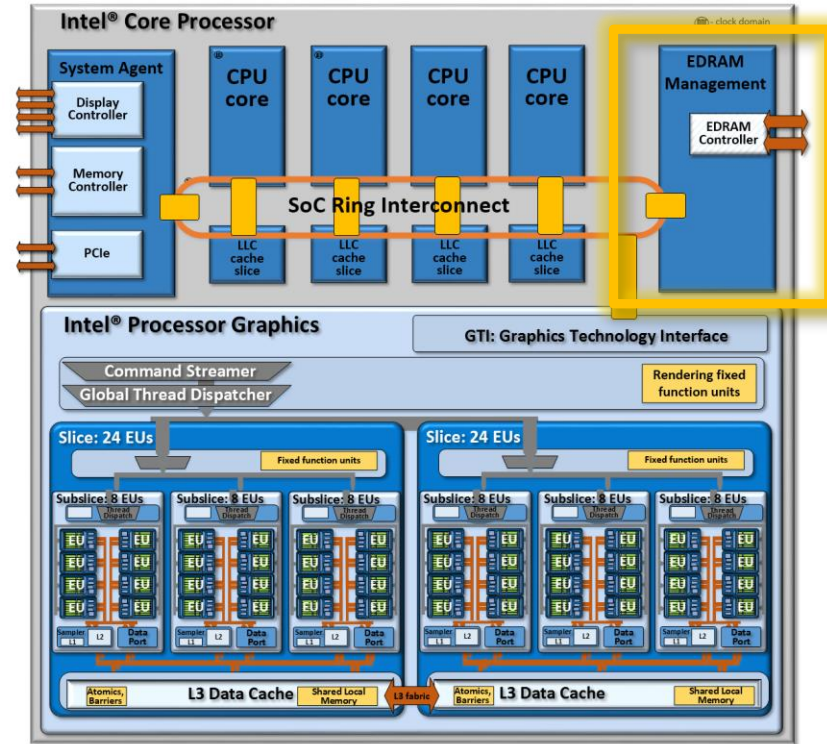
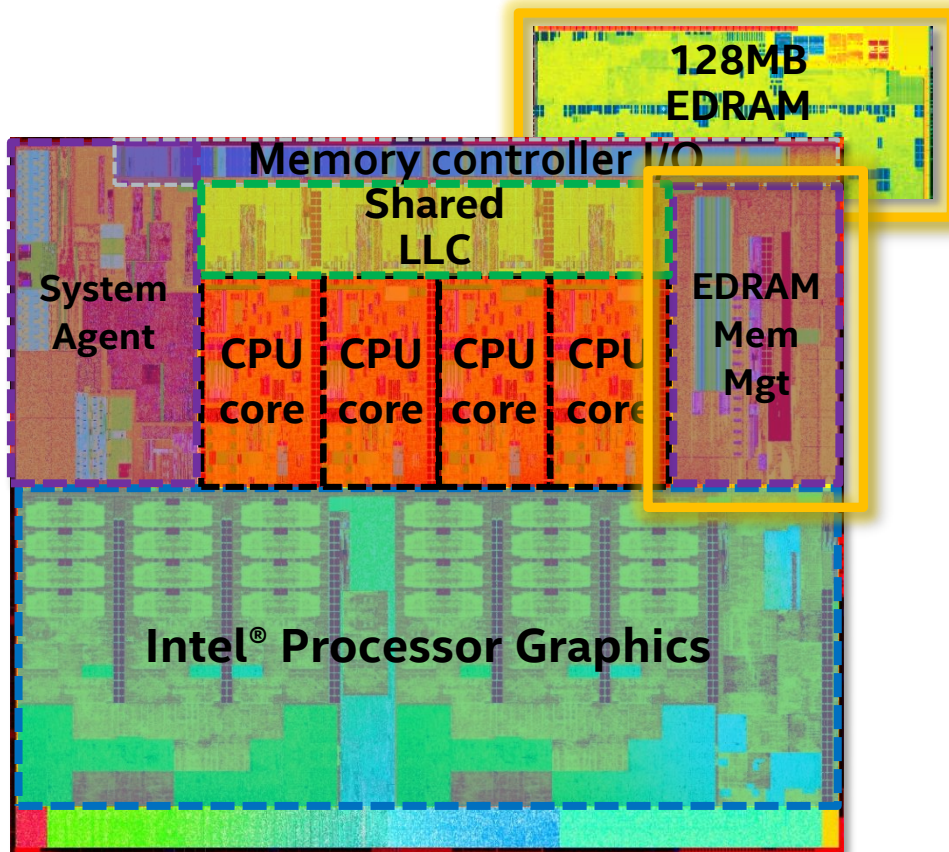
Intel® Core™ M: 4-5 Watts Enabling Fanless Laptops



- Many different processor products, with different processor graphics configs
- Multiple CPU cores, shared LLC, system agent
- Multiple clock domains, target power where it's needed

Note: Intel 5th Generation Core™ example shown

4 Cores & Iris Pro: Powering Performance PC



Note: Intel 5th Generation Core™ example shown

Complimentary Computing Engines

ILP

DLP

TLP

CPU: General Purpose

ILP

DLP

TLP

GPU: Parallel Data Crunching

ILP: Instruction Level Parallelism; **DLP**: Data Level Parallelism; **TLP**: Thread Level Parallelism

Complimentary Computing Engines

ILP

- Superscaler
- Out-of-order
- Branch prediction

DLP

- SIMD

TLP

- Hyper-threading
- Multi-Core

CPU: General Purpose

ILP

- Multi-issue

DLP

- SIMD
- SIMP

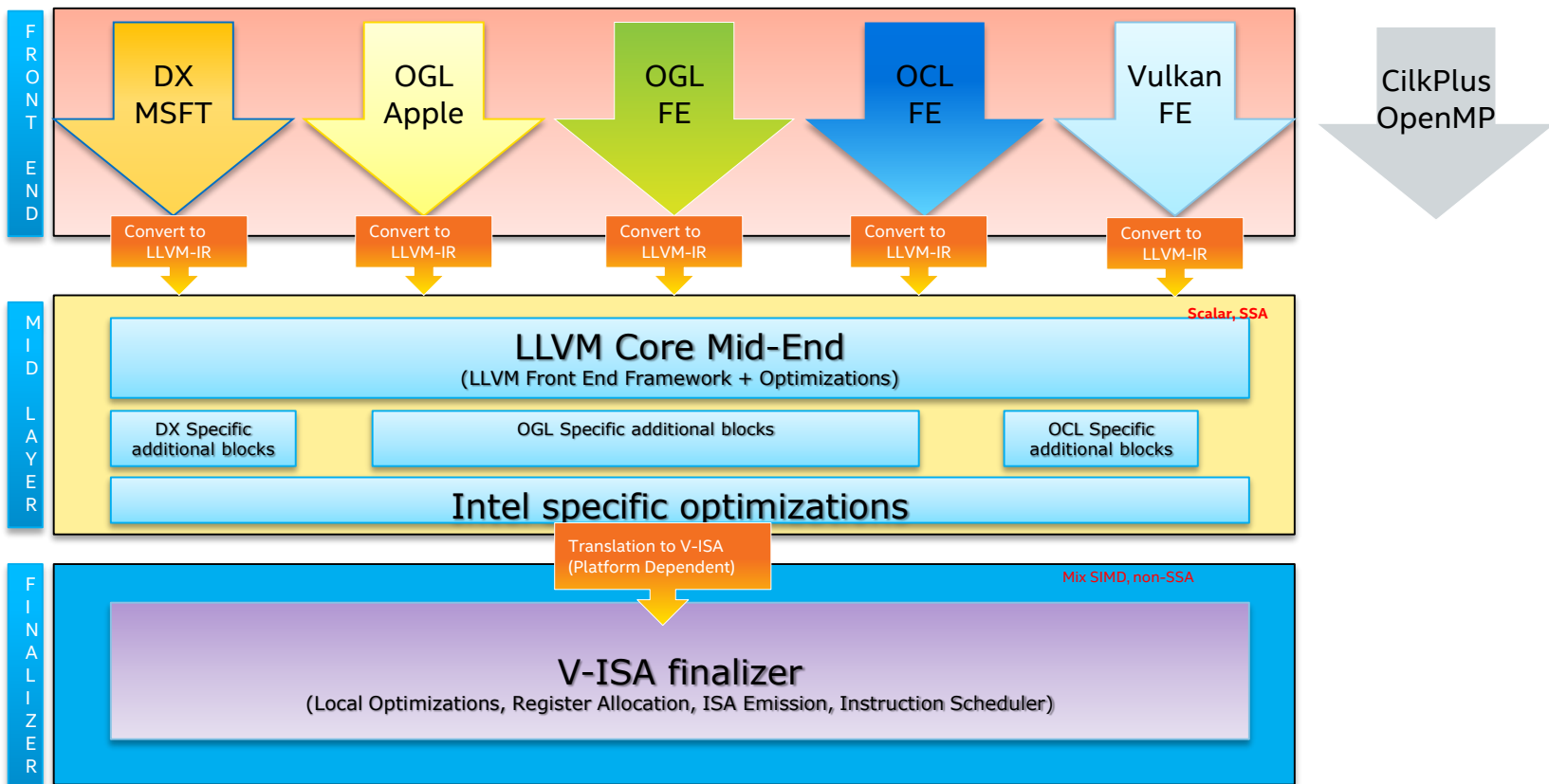
TLP

- Multi-thread
- Many Cores

GPU: Parallel Data Crunching

ILP: Instruction Level Parallelism; **DLP**: Data Level Parallelism; **TLP**: Thread Level Parallelism

GPU Programming Mostly through Device API & Languages



Intel Graphics Compiler Architecture

Taking Advantages of both CPU's and GPU's

Opportunities

- Most devices already have both CPU and GPU
- Meeting compute demand with limited power
 - 3D Graphics
 - Multi-Media
 - Imaging
 - Computational Photography
 - Computer Vision (e.g. OpenCV/VX)
 - Deep Neural Network (e.g. Caffe)

Challenges

- Parallel programming is hard
 - (use vTune tool suite)
- Device programming is hard
 - (use vTune tool suite)
- Impact device responsiveness (GPU drives the screen)

BACK UP

Intel® Processor Graphics

- These details and more available in our architecture whitepapers:

<https://software.intel.com/en-us/articles/intel-graphics-developers-guides>



**Whitepaper:
The Compute Architecture of
Intel Processor Graphics Gen8**



**Whitepaper:
The Compute Architecture of
Intel Processor Graphics Gen9**

Read our whitepapers

Legal Notices and Disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.
- No computer system can be absolutely secure.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit <http://www.intel.com/performance>.
- Intel, the Intel logo and others are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.
- © 2015 Intel Corporation.

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

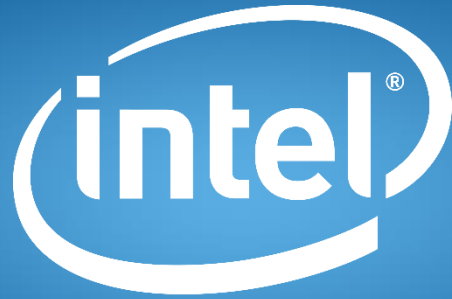
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Copyright © 2015, Intel Corporation. All rights reserved. Intel, Pentium, Xeon, Xeon Phi, Core, VTune, Cilk, and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804



experience
what's inside™

Algorithms vs. Architectures: Opportunities and Challenges in Multicore/GPU DSP

Nick Moore

16 December 2015

Algorithms vs. Architectures vs. Tools: Opportunities and Challenges in Multicore/GPU DSP

Nick Moore

16 December 2015

Opportunity: Tools for Concurrent Systems

- Concurrency in hardware is not new, but still remains a challenge
 - Herb Sutter timeline: “The Free Lunch Is Over” ~2005 & “Welcome to the Jungle” ~2011
- Architecture innovations have pushed performance boundaries
 - Requires new algorithms (hard)
 - Develop on heterogeneous and/or concurrent hardware (hard)
- How do we allow people to focus on their goals?
- Of interest: granularity and constraints in programming abstractions

Granularity

- Systems/applications as collections of algorithmic building blocks
 - Needed for modularity, reuse, and development scalability
 - Looking at different programming paradigms for each
- Systems: declarative styles offer benefits for concurrency
 - Easier adoption than at the algorithm level
 - Graphical UI for design entry has many strengths in this context
 - Architecture and schedule agnostic actor model and dataflow (data-driven) execution
 - Express computation without primitives that cause trouble

Constraints

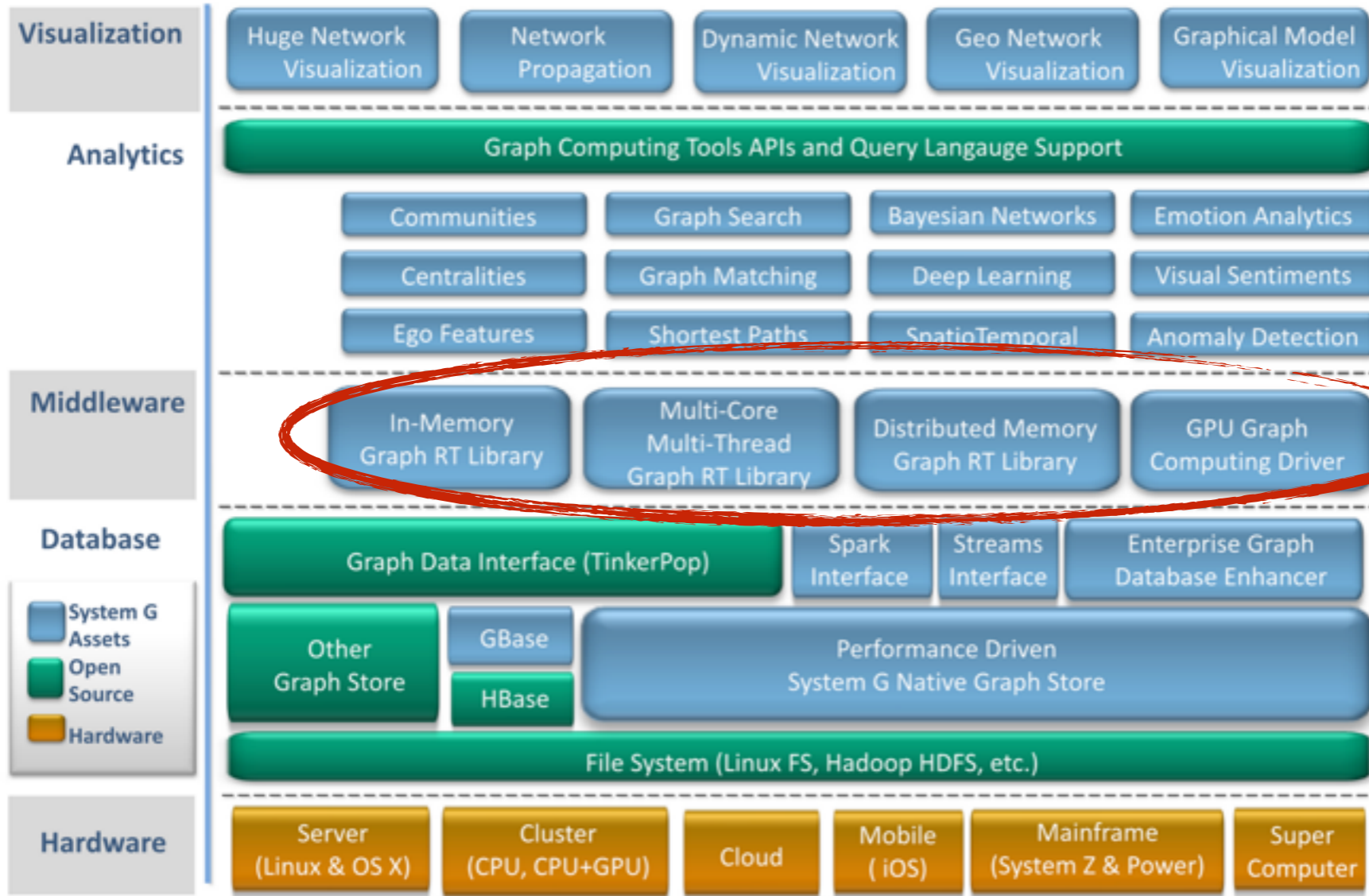
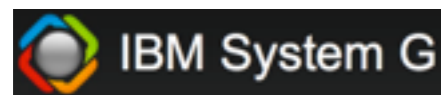
- Efficiency from general abstractions is hard
 - Many fundamentally different types of applications
 - Flexible and robust techniques for mapping applications to diverse architectures
 - Optimization across levels of granularity and algorithmic building blocks
 - Two scenarios: desktop simulation/development and deployment
- Reexamine constraints: need to be explicit about all constraints chosen
 - Domain: streaming signal processing, communications, and computer vision
 - Programming models that impose desired constraints but work for the domain
 - Balancing constraints against broad usefulness
 - Continue to educate users about the new reality – and give them good tools

Algorithms vs. Architectures: Opportunities and Challenges in Multicore/GPU DSP

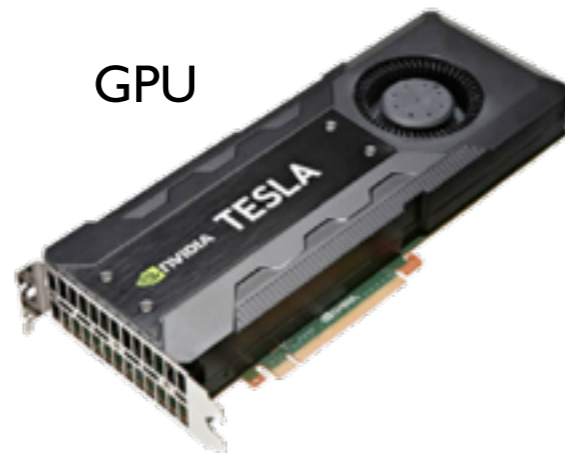
Yinglong Xia

IBM T.J. Watson Research Center

Architectural Diversity



GPU

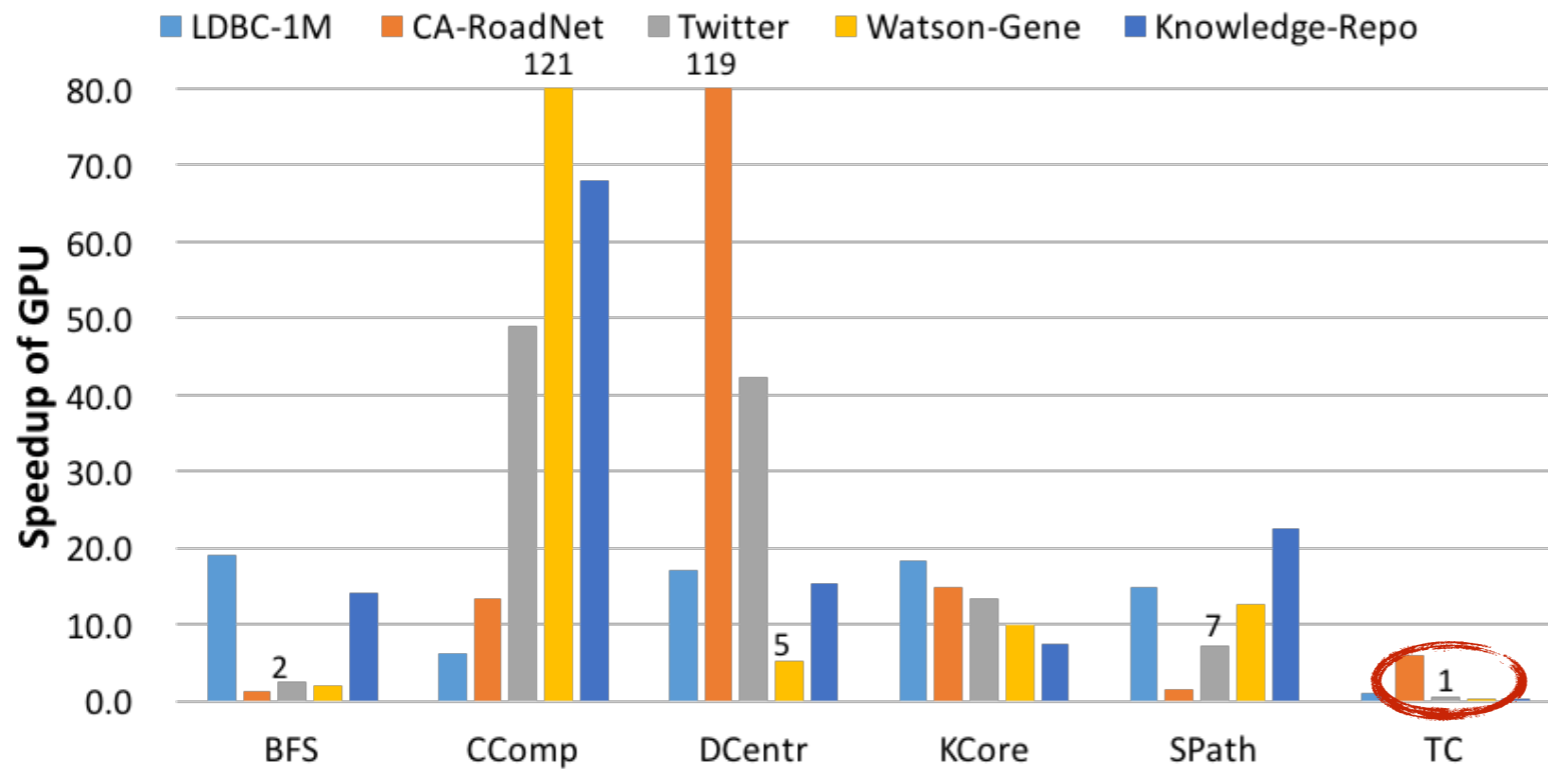


Mobile Processor



P8 Cluster

Challenges in Obtaining Performance



Multicore v.s. GPU

GraphBIG@github:
<https://github.com/graphbig>

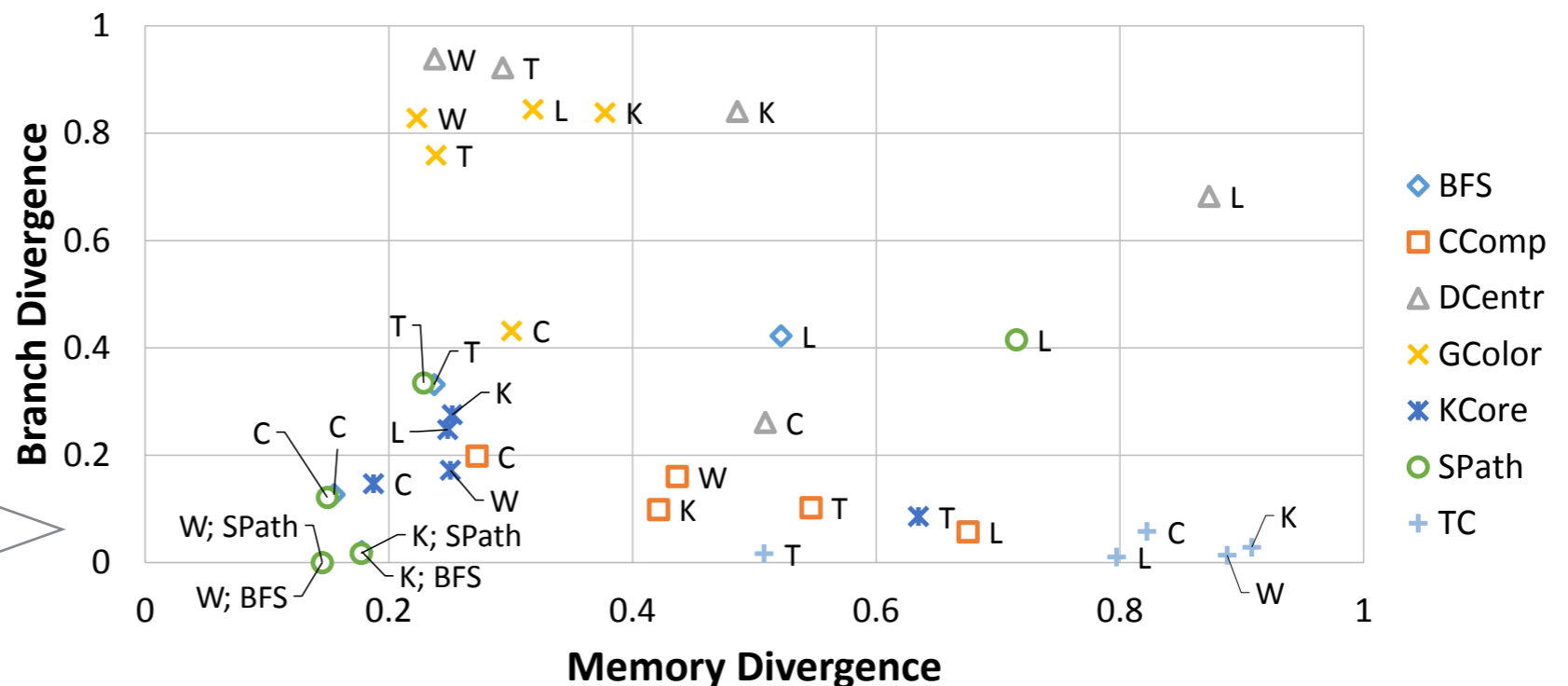


Given a set of input graphs, a set of graph analytics, and two parallel computing architecture (Multicore, GPU), the obtained performance is not consistent.

Co-design:

- Algorithm
- Architecture
- Data

Analyze the GPU performance for graph computing by looking into the architectural divergence.



L: LDBC-1M

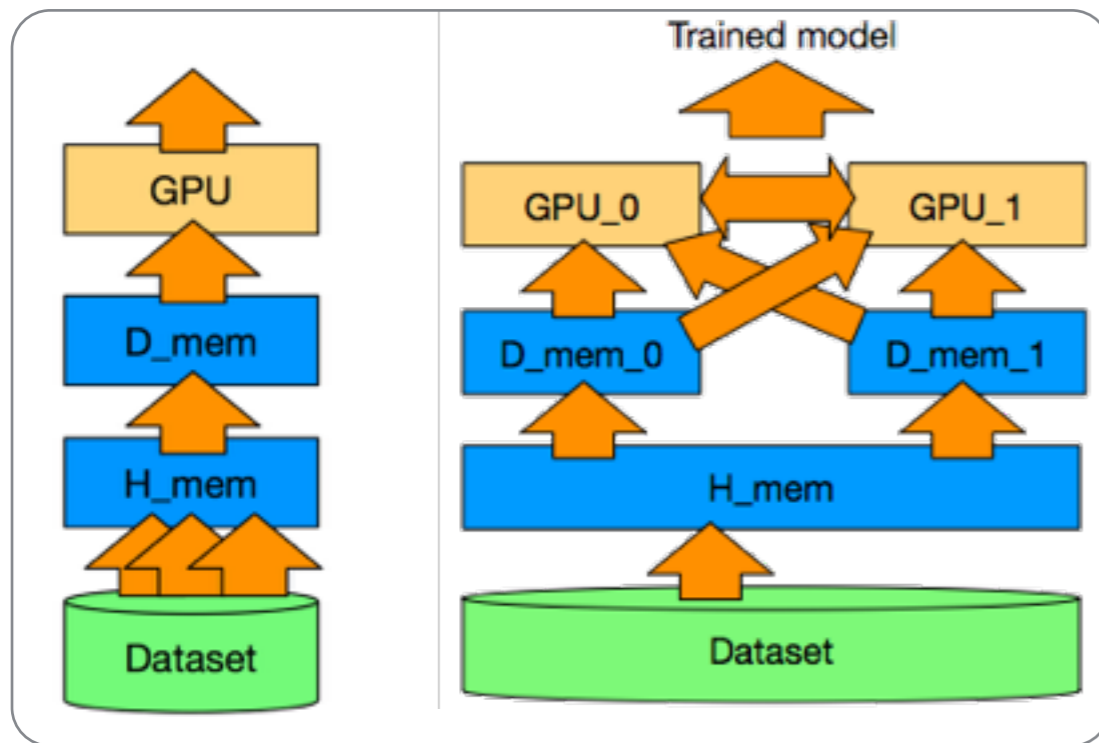
C: CA-RoadNet

T: Twitter

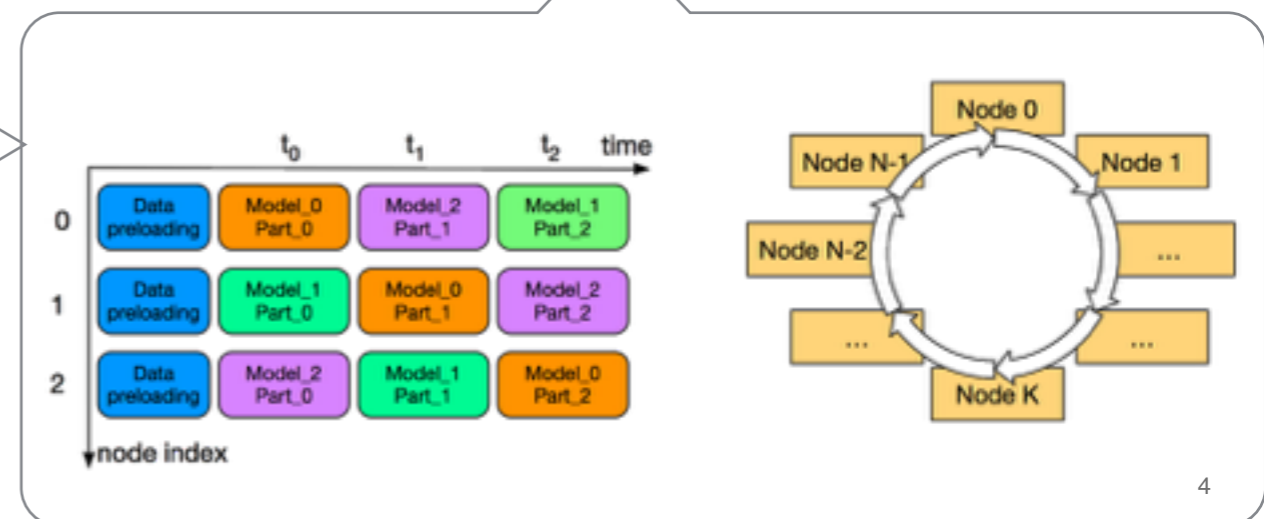
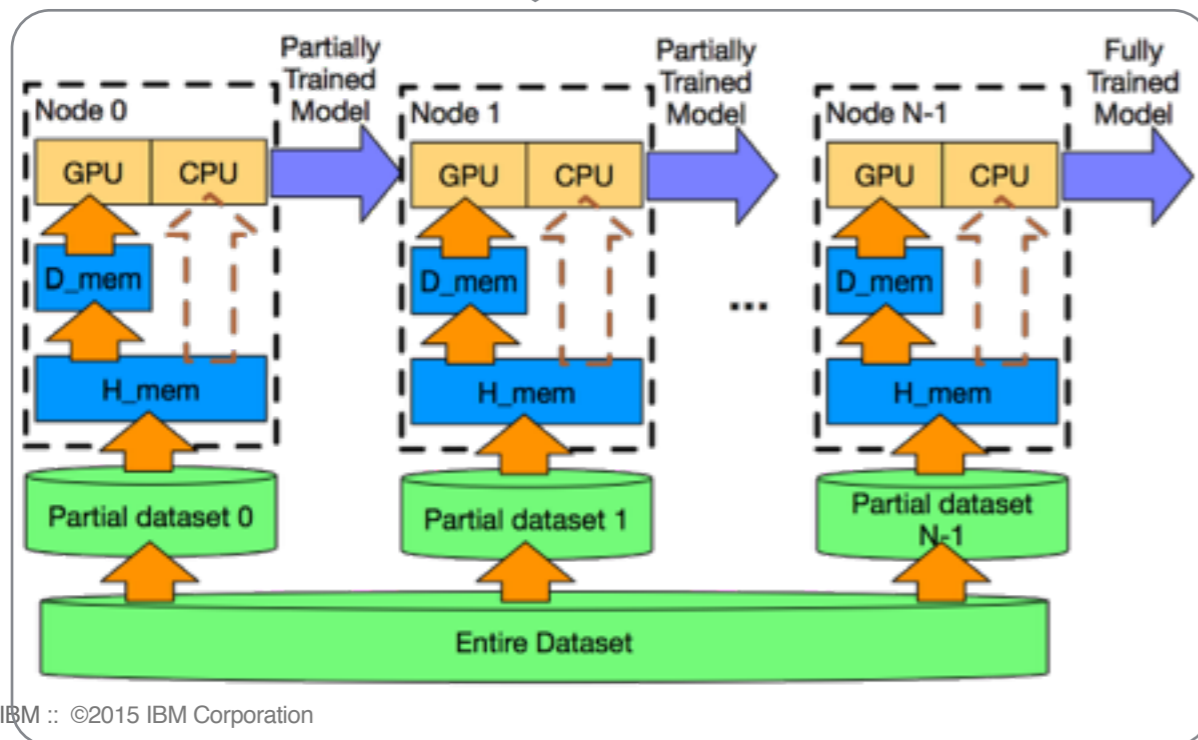
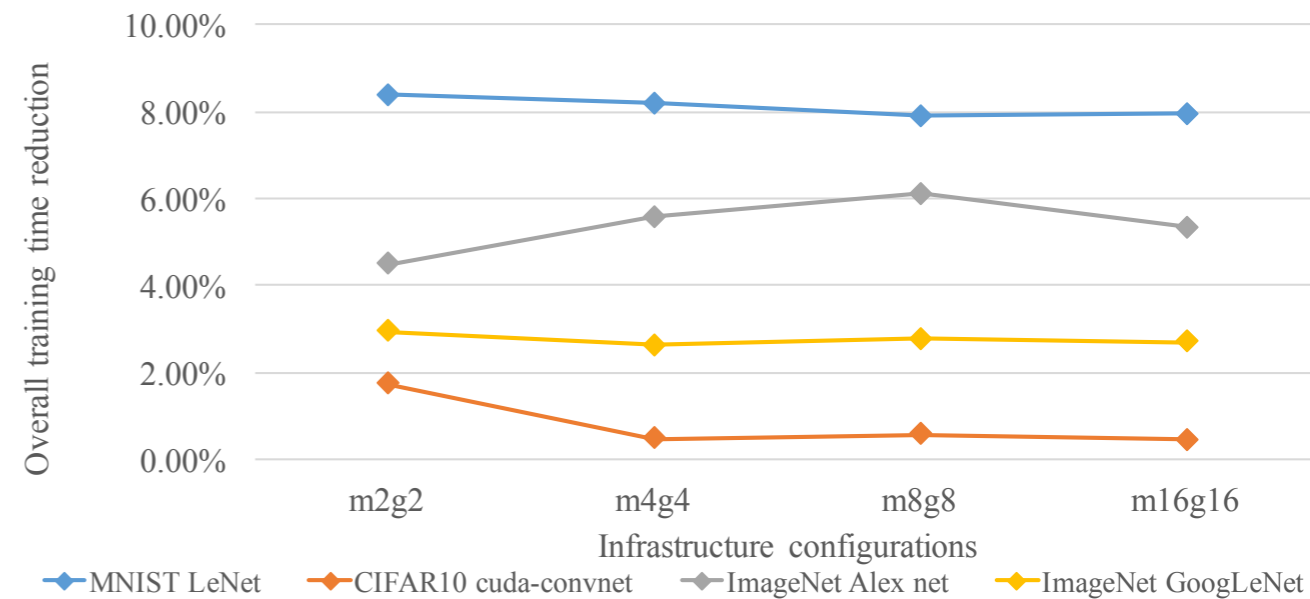
W: Watson-Gene

K: Knowledge-Repo

An Example of Deep Learning on CPU-GPU Cluster



Cognitive computing cluster	
Number of nodes	160
Network connection	10 Gb/s Ethernet and 56 Gb/s InfiniBand
Hardware per node	
CPU	1 Intel(R) Xeon(R) E5-2667 v2 @ 2.70GHz 8 cores, 2 threads/core with 264 GB memory
GPU	2 NVIDIA Tesla k40m with 12 GB RAM per GPU



Some Co-Design Considerations

- Unified specification of heterogeneous systems
- Reasonable Hardware/software partitioning
- Scalable scheduler at runtime
- Performance modeling and algorithm mapping
- ...

Thanks

Deployed
STories

Also mentioned
in SQA

139

EL Enablement
Material

(pre-deployment)

135

New Expense

landing page