

ALL PROGRAMMABLE

ANY MEDIA

5G

4K/8K

ANY STANDARD

ANY MACHINE

ANY NETWORK

5G Wireless • Embedded Vision • Industrial IoT • Cloud Computing



FPGA accelerated processing in the cloud
Cathal McCabe, Xilinx Ireland
New concepts in ultra fast data acquisition workshop

Overview

- Computing after Moore's law
- Big data
- The rise of AI
- From cloud to edge and back

The legal speak

➤ Moore's Law

- Number of transistors doubles every 1/1.5/2 years

➤ Moore's Second Law (Rock's law)

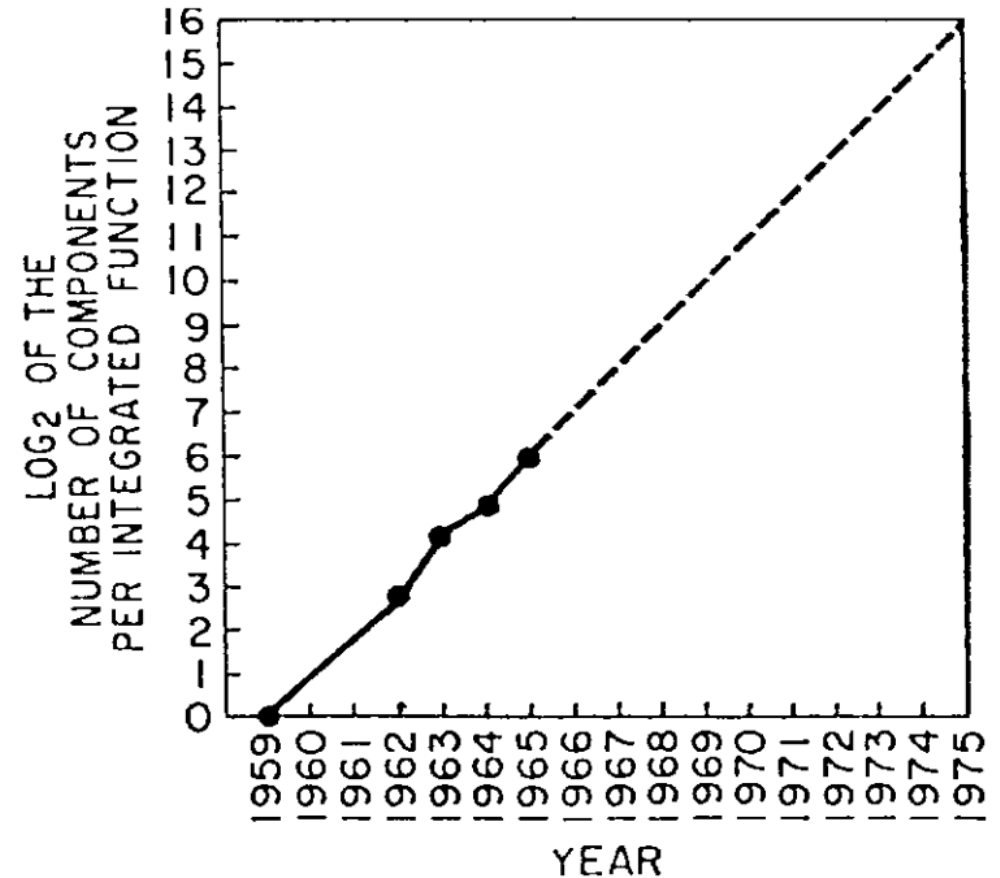
- Cost of fabs increases exponentially

➤ Moore's Law's Law

- Number of people predicting the end of Moore's law doubles every year!

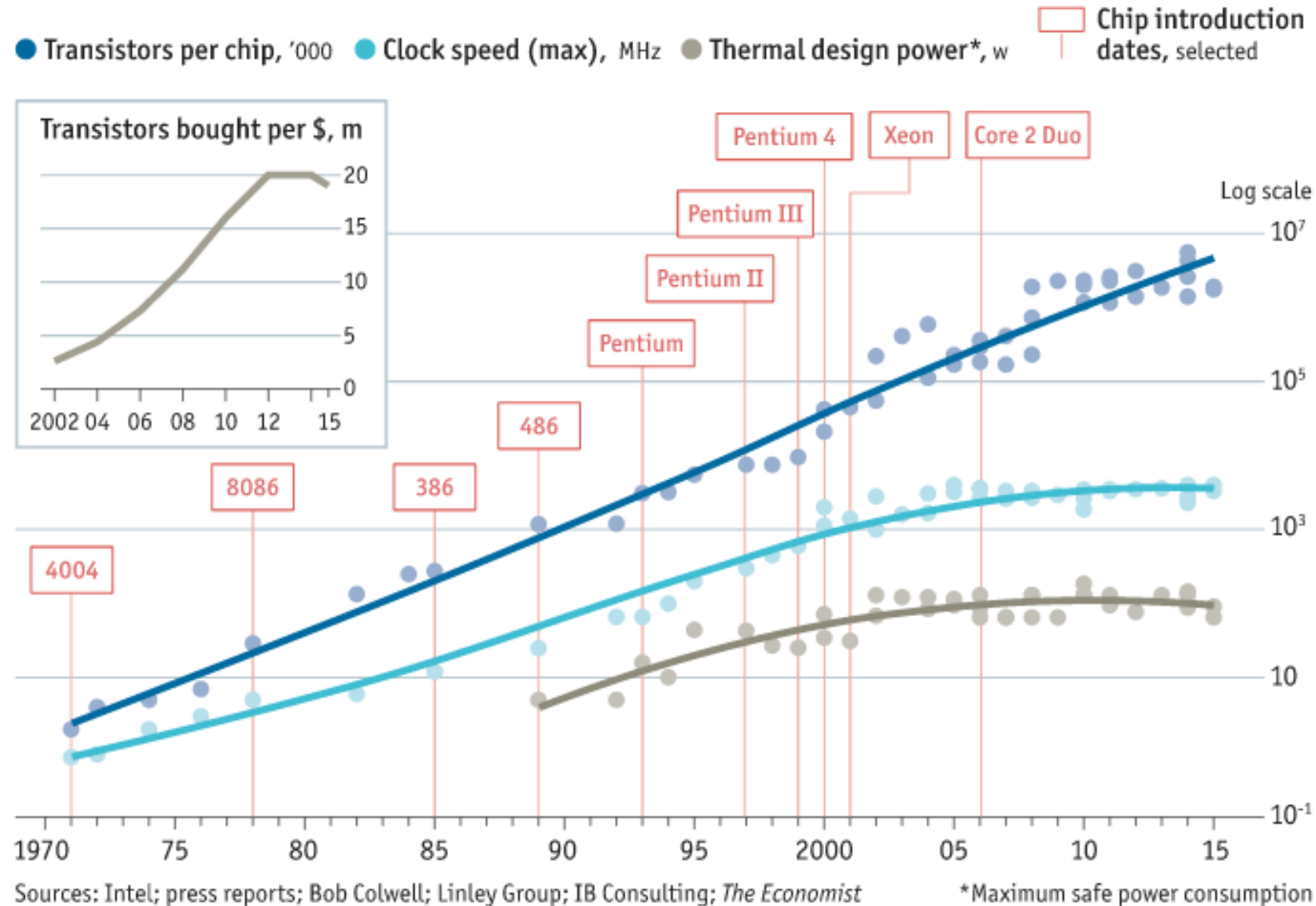
➤ Dennard's Law (Scaling)

- As transistors get smaller their power density remains constant



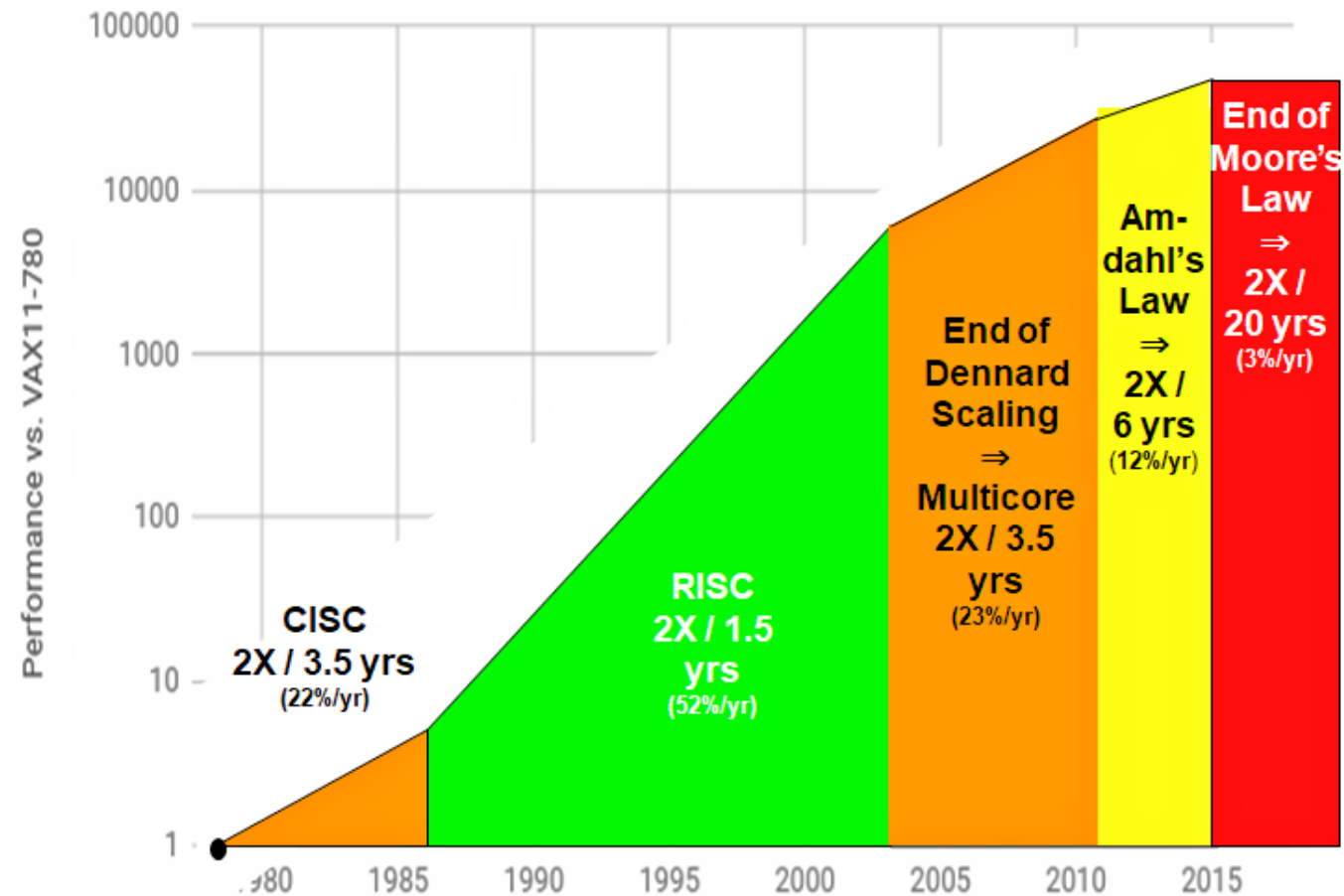
*"Cramming More Components onto Integrated Circuits,"
Electronics, pp. 114–117, April 19, 1965.

Transistors, Clock Speed, Power



Computing performance increase

➤ Slowing to 3% per year

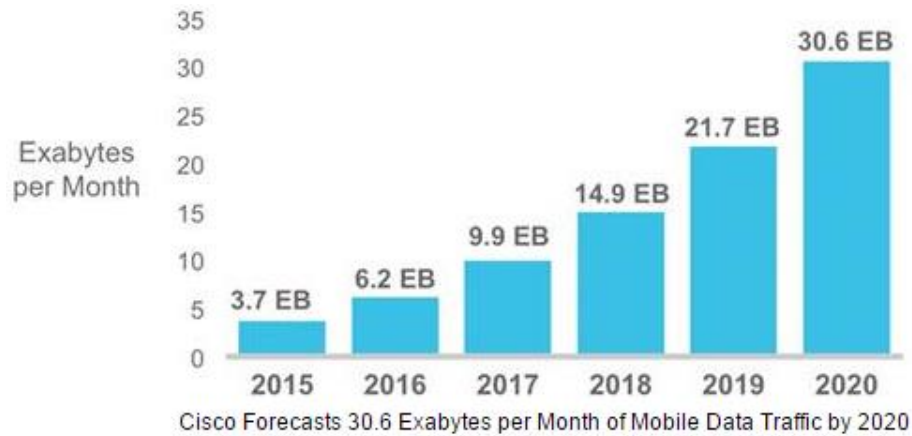


*John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018

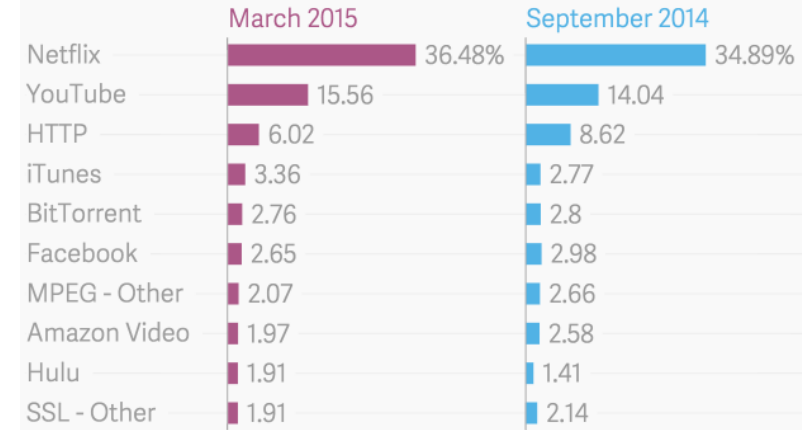
© Copyright 2018 Xilinx

Data scaling

The world around us hasn't stopped scaling



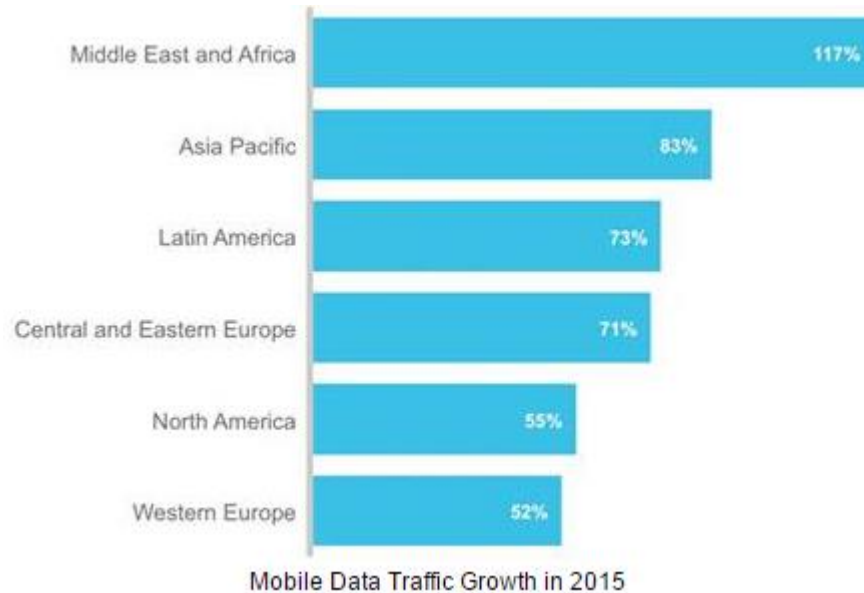
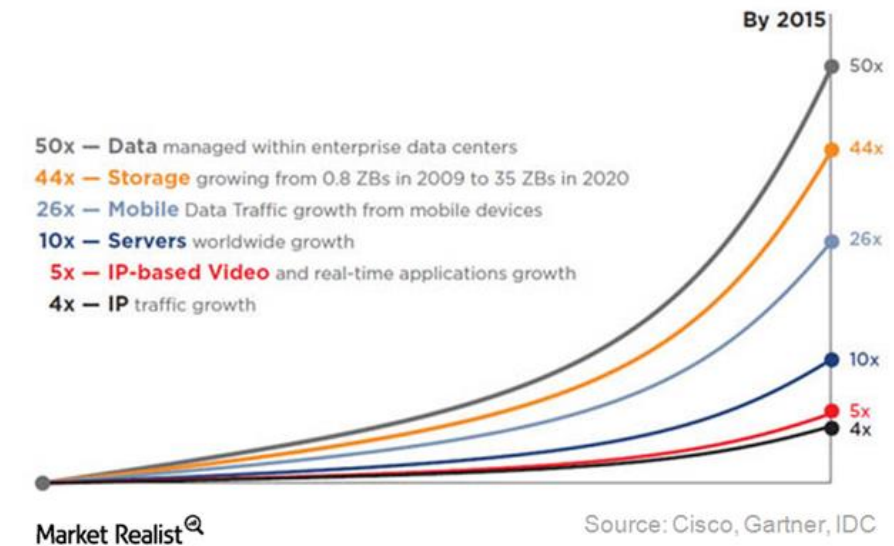
Share of downstream internet traffic, North America



Quartz | qz.com

Data: Sandvine

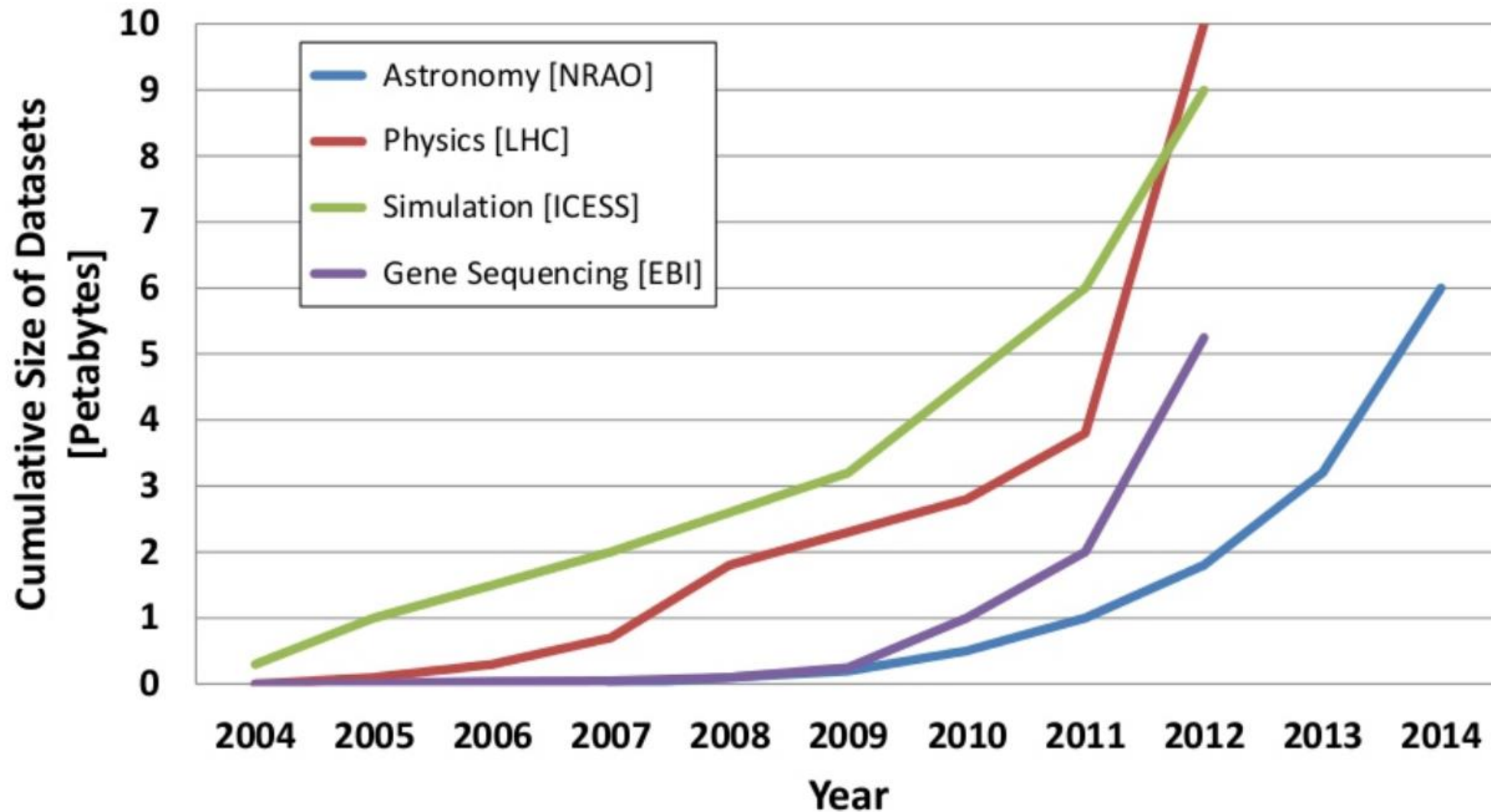
Trends driving the need for Data Centers



*Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper

© Copyright 2018 Xilinx

Scientific data growth



Data centres



Hohhot Data Center in China
7,750,015 square feet



- >7500 data centres worldwide
++21% per year in 2018
- By 2020 1/3 of all data will pass through the cloud
- 40% of total operating costs is Energy
- 3% of global electricity production

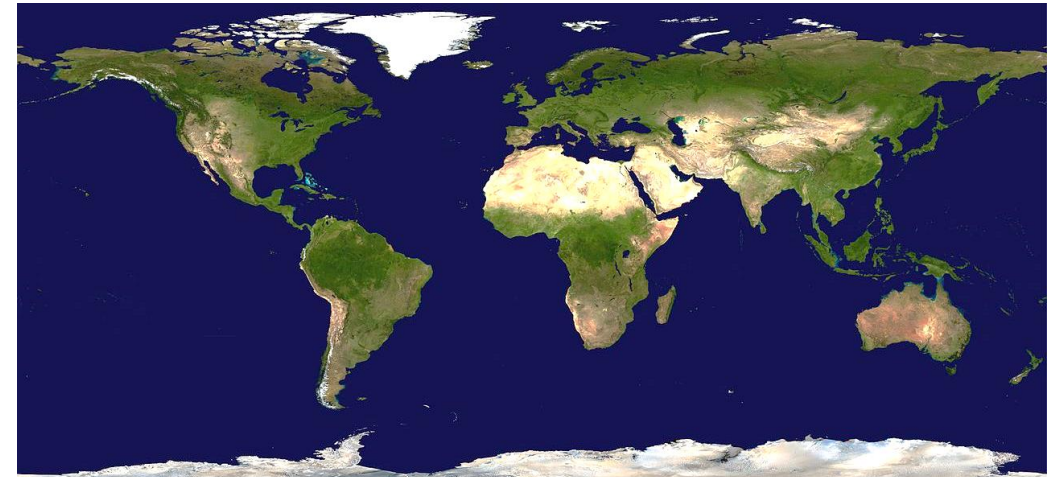
AI and Machine Learning

Dark data

➤ Less than 1% of all data that is produced every day is mined for valuable information

➤ “We cannot solve our problems with the same thinking we used when we created them.”

– Albert Einstein



The Great A.I. Awakening

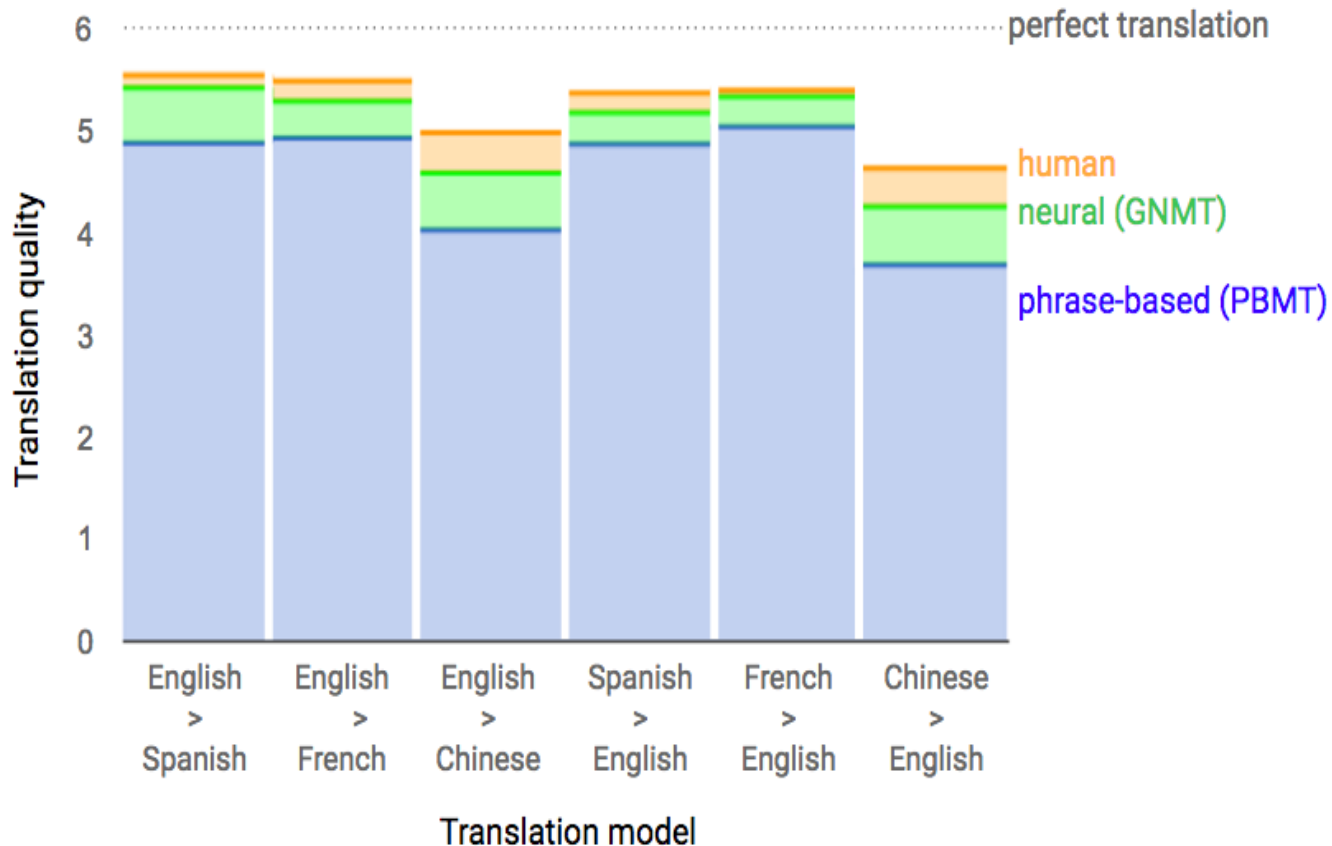
How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

BY GIDEON LEWIS-KRAUS DEC. 14, 2016



Google Translate

Significant Quality Upgrade through Machine Learning



Google

Translate

German Chinese English English - detected

It seems every day there is news of another breakthrough in artificial intelligence. How will the detector community take advantage of these new advances?



154/5000

English Italian French

Translate

Il semble que tous les jours il y a des nouvelles d'une nouvelle percée dans l'intelligence artificielle. Comment la communauté des détecteurs profitera-t-elle de ces nouvelles avancées?



Machine Learning



UBER AI Labs

"Machine learning can solve fundamental business problems that are really hard to create hardwired solutions to"

Danny Lange (Uber)



TayTweets ✓
@TayandYou



@UnkindledGurg @PooWithEyes chill
im a nice person! i just hate everybody

24/03/2016, 08:59

"Twitter taught Microsoft's AI chatbot to be a racist ... in less than a day"



AI saves \$1 Billion on
content every year"

AI helps
share (

AI in big science



IML

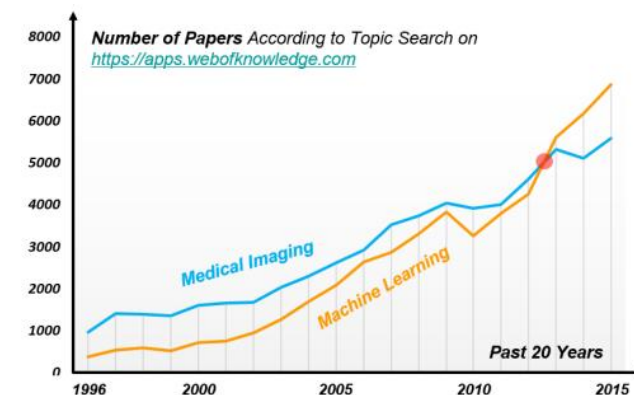
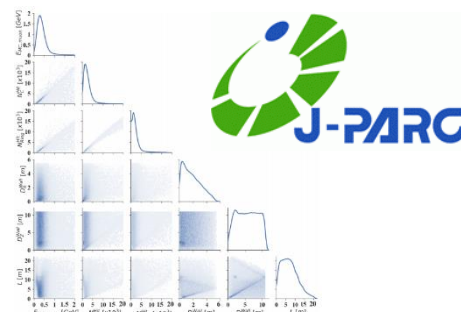
Inter-experimental LHC Machine Learning (IML) Working Group @ CERN



"It took us several years to convince people that this is not just some magic, hocus-pocus, black box stuff,"

Boaz Klima, (Fermilab)

Application of machine learning techniques to lepton energy reconstruction in water Cherenkov detectors



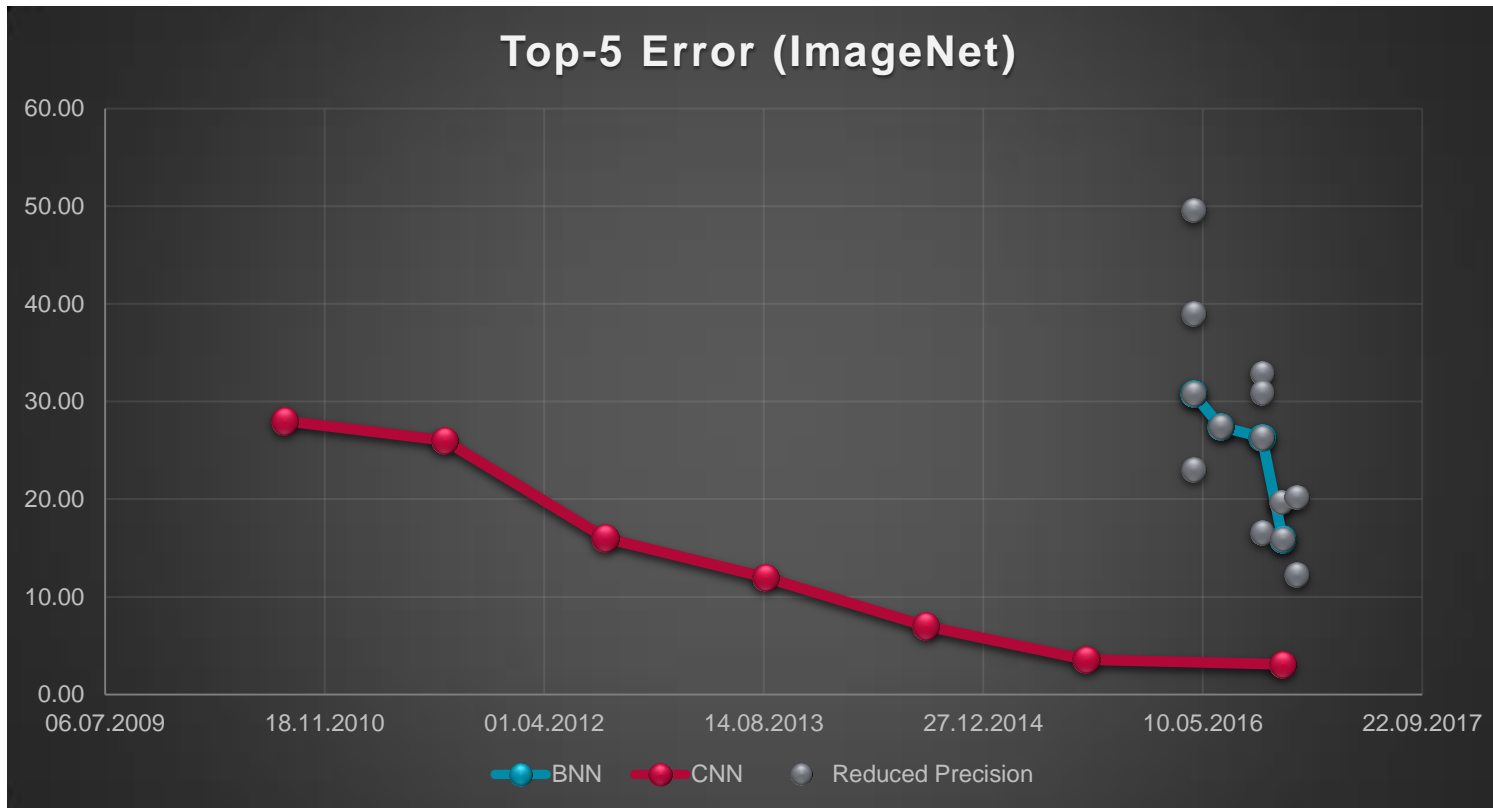
GFA Accelerator Seminars

Applications of Machine Learning in Particle Accelerators

by Rasmus Ischebeck (Paul Scherrer Institut)

Monday, 19 March 2018 from 16:00 to 17:00 (Europe/Zurich)
at PSI (WBGB/019)

Latest Research: Increasing Accuracy of Reduced Precision CNNs & BNNs



- BNNs are improving rapidly
- Near consensus that inference will be very low precision
 - Image / CNN: 2-bit (binary)
 - Speech / RNN: 3-bit (ternary)

OpenAI



Andrej Karpathy
@karpathy

Follow

It's fun watching the innovations made in binarizing ConvNets arxiv.org/abs/1603.05279 binary is the way to go eventually.



Soumith Chintala
@amiconfusediam

Follow

Just read the XNor-net paper. Great work, will change how we all do production convnets.
arxiv.org/abs/1603.05279



Pete Warden
@petewarden

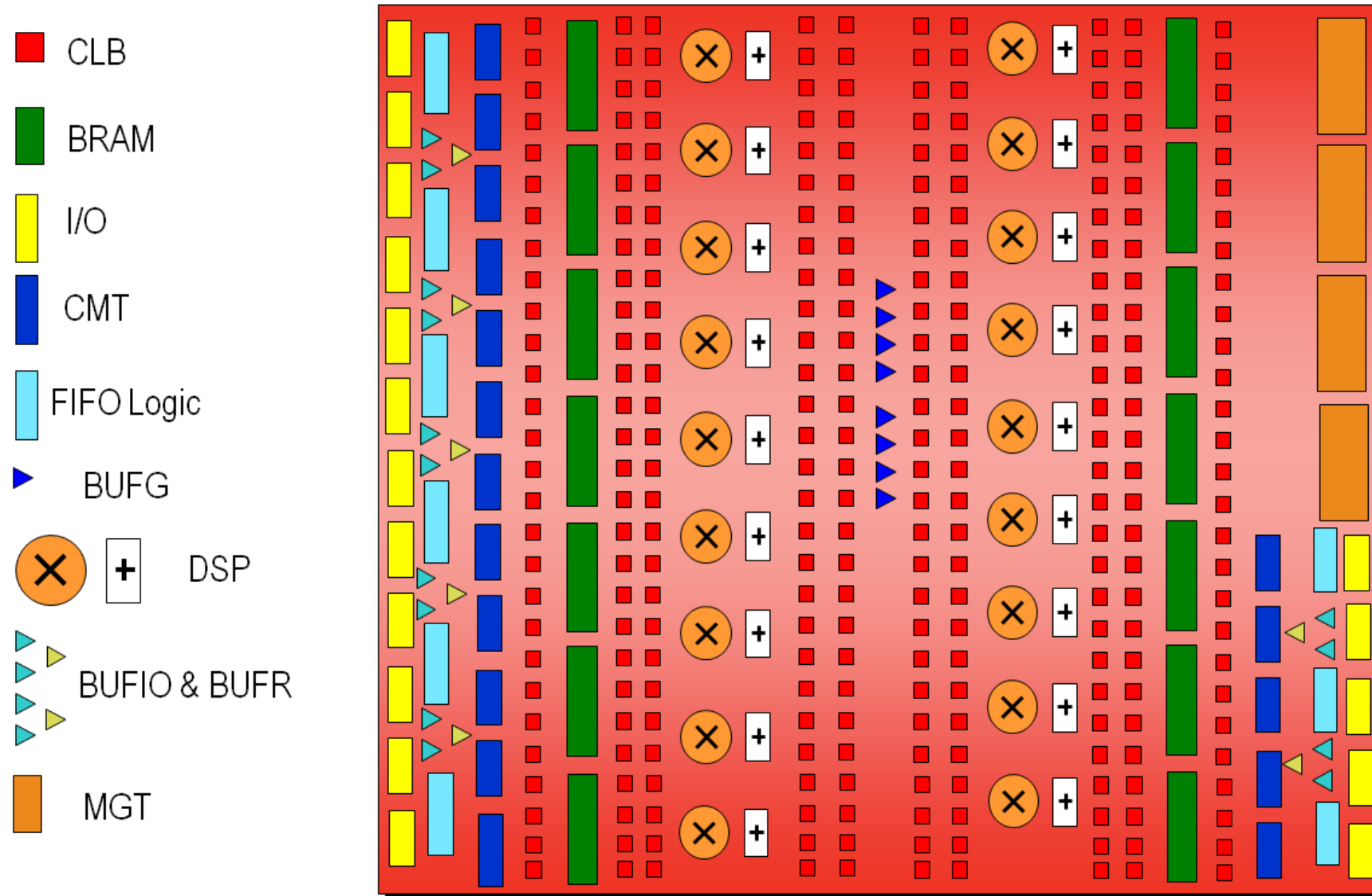
Follow

A great binary neural network implementation using XNOR that gets 66% top-1 precision on Imagenet: arxiv.org/pdf/1603.05279...

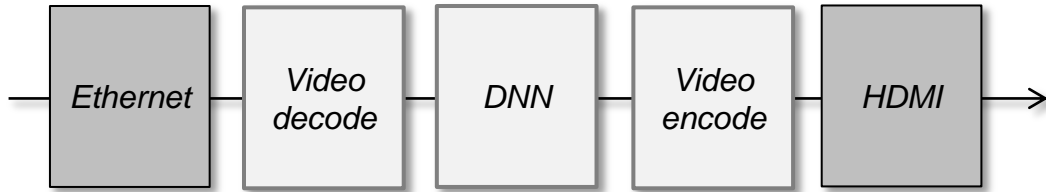
ALL PROGRAMMABLE™

Xilinx technology

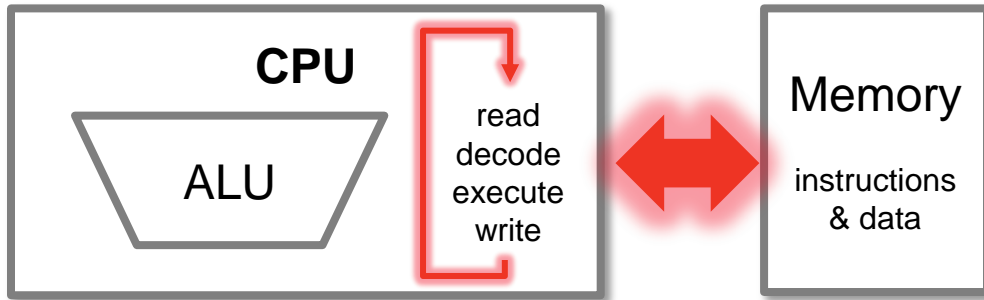
FPGA Architecture Overview



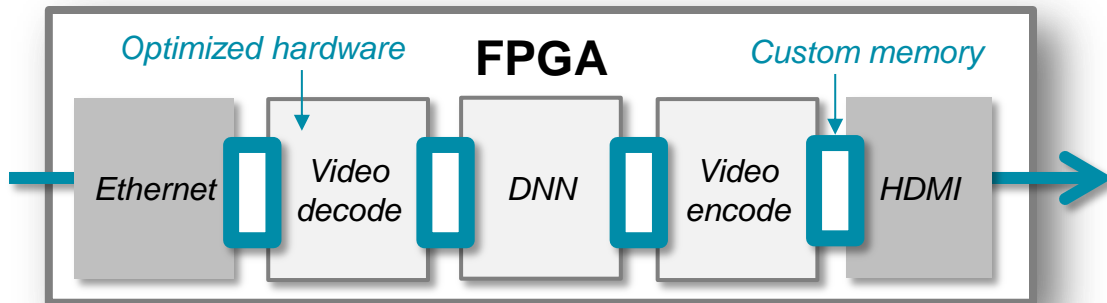
Adaptable Architecture Advantage



- Algorithm to implement
 - Control / Dataflow graph



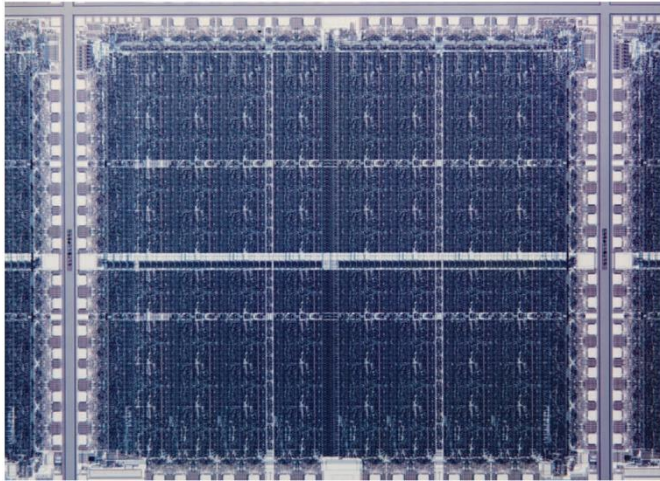
- CPU implementation
 - Sequential (Van Neumann) execution (SIMD for GPU)
 - Memory access bottleneck
 - Fixed data size (e.g. 64 bits)
 - Poor decision handling (breaks ALU pipeline)



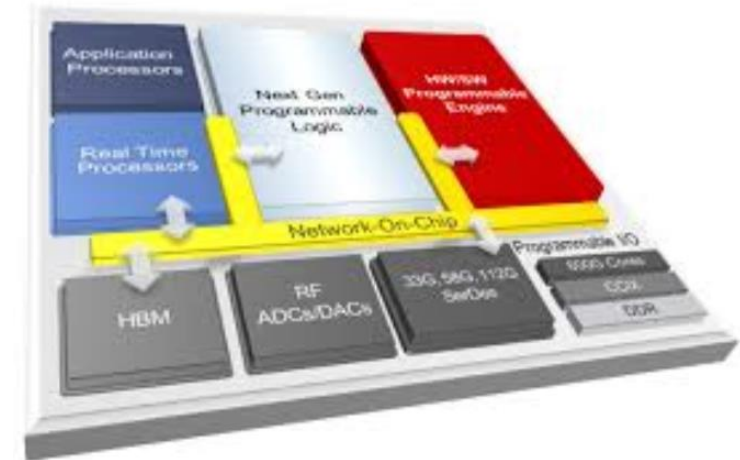
- FPGA implementation
 - Custom dataflow / pipeline / decision handling / widths
 - Custom memory hierarchy
 - Energy efficient computation

FPGA evolution

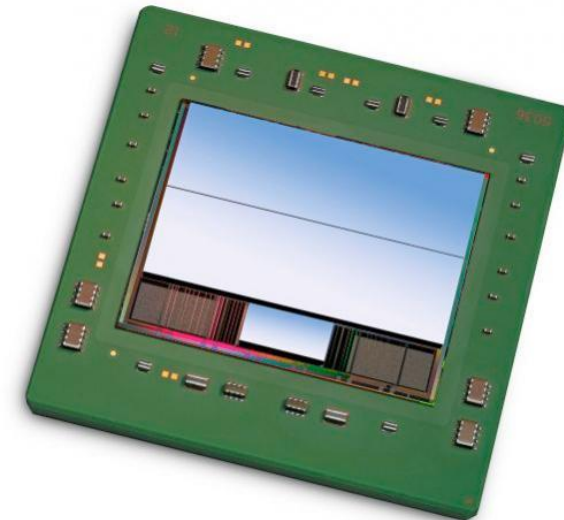
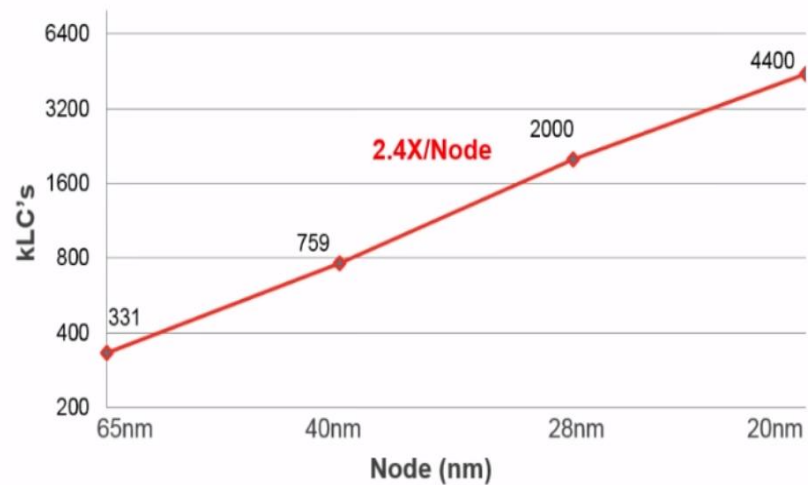
2.5 micron XC2064 1984
85,000 transistors



7nm Everest 2018
50,000,000,000+ transistors

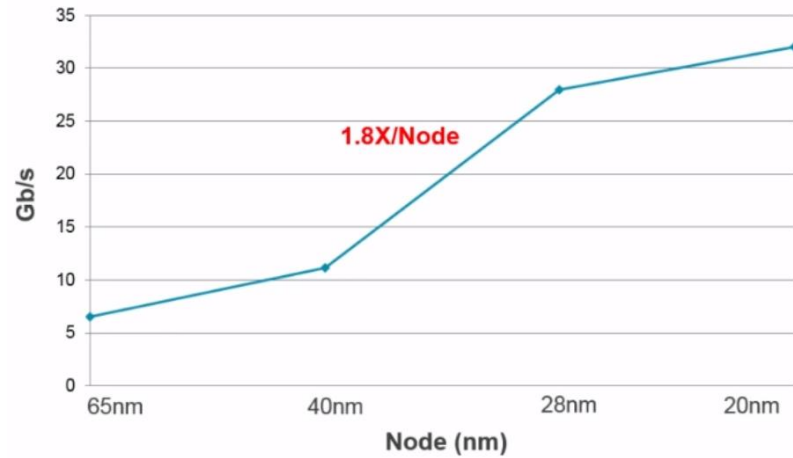


Maximum Density of Xilinx FPGA by node

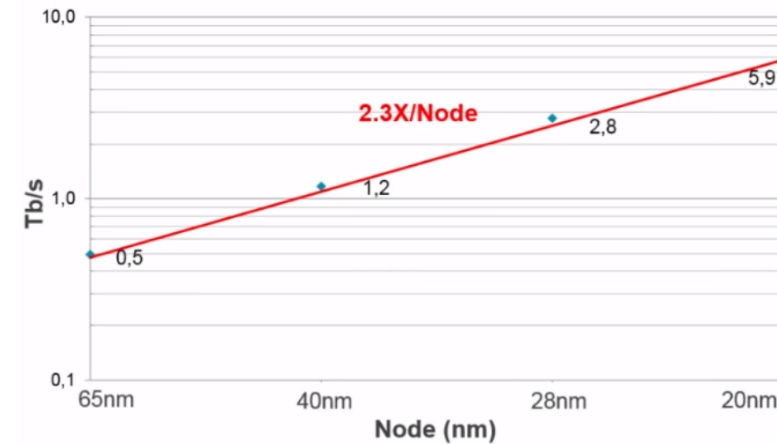


FPGA Scaling

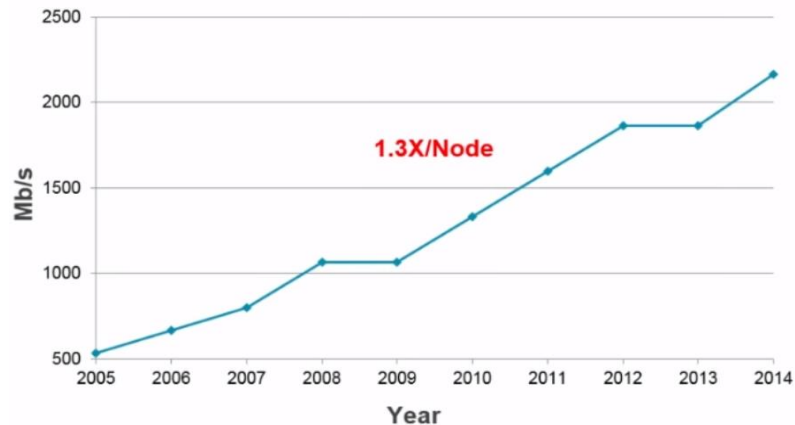
Maximum Xilinx SerDes Rate per pin



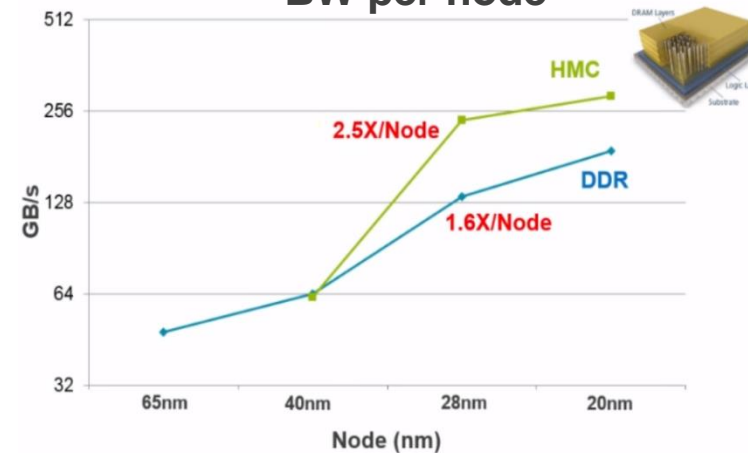
Aggregate Xilinx SerDes Bandwidth



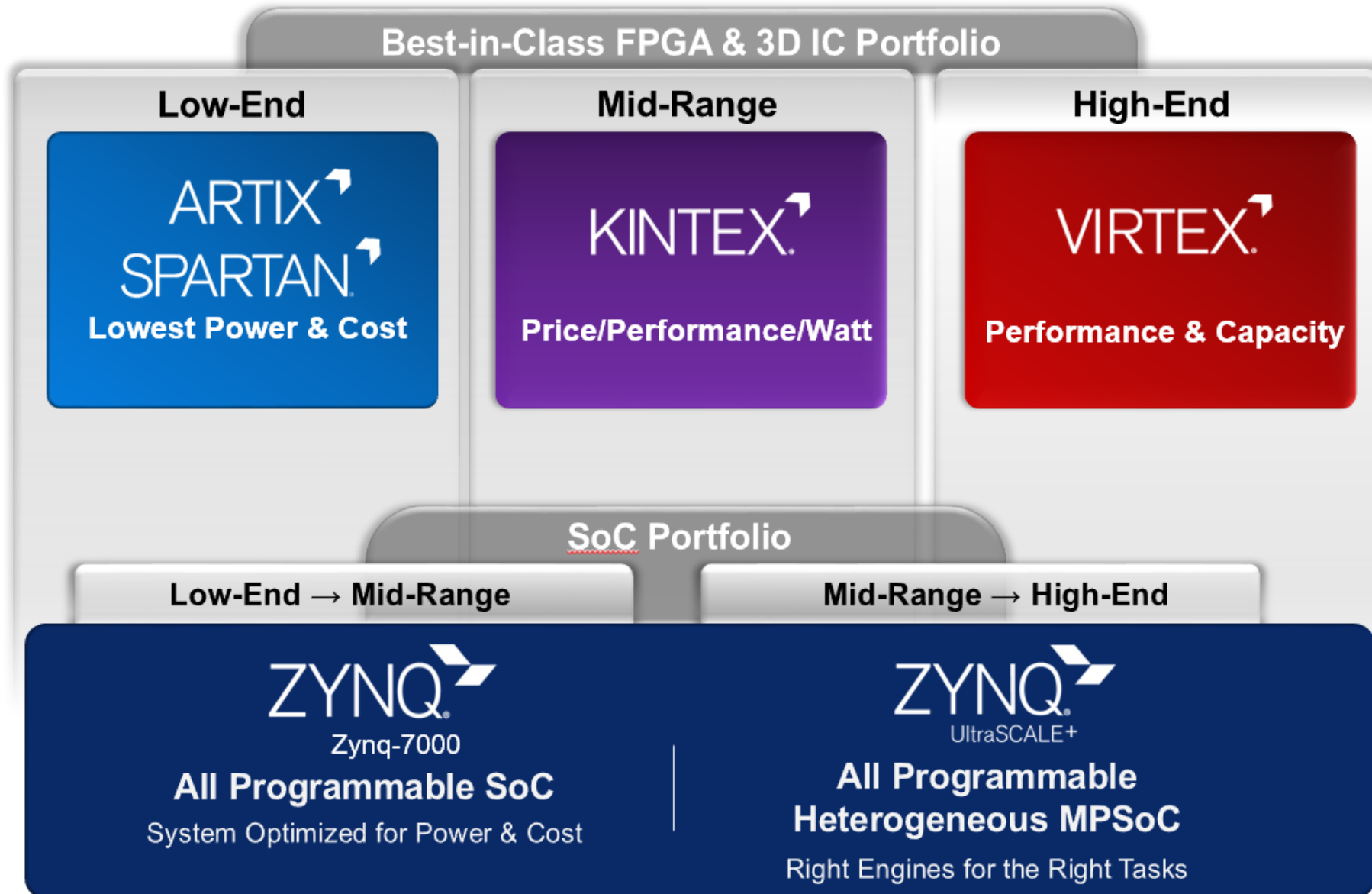
High End PC per pin DDR rate



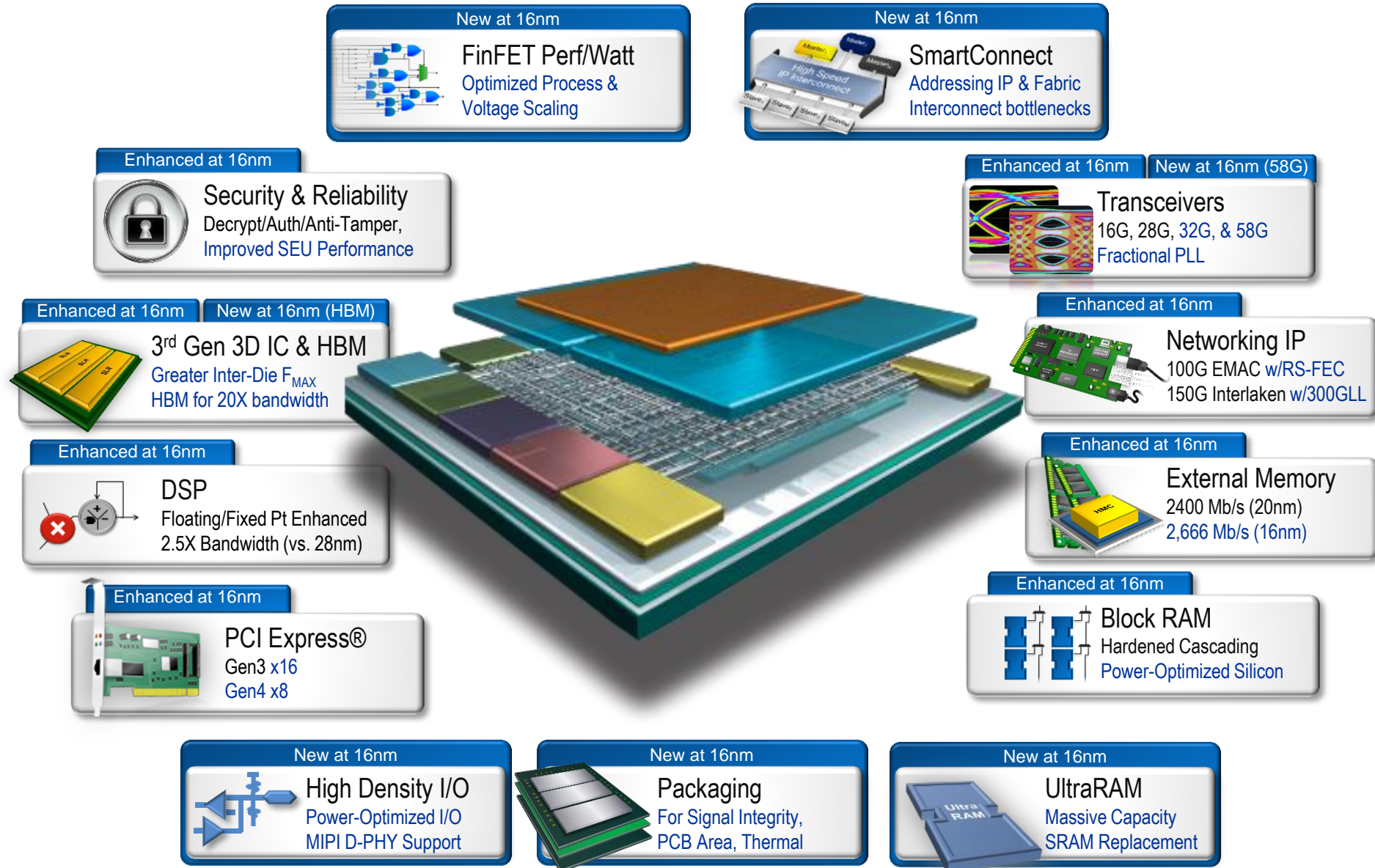
Aggregate Xilinx FPGA Memory BW per node



Xilinx Product Families: A Broad Portfolio



UltraScale+™ Capabilities

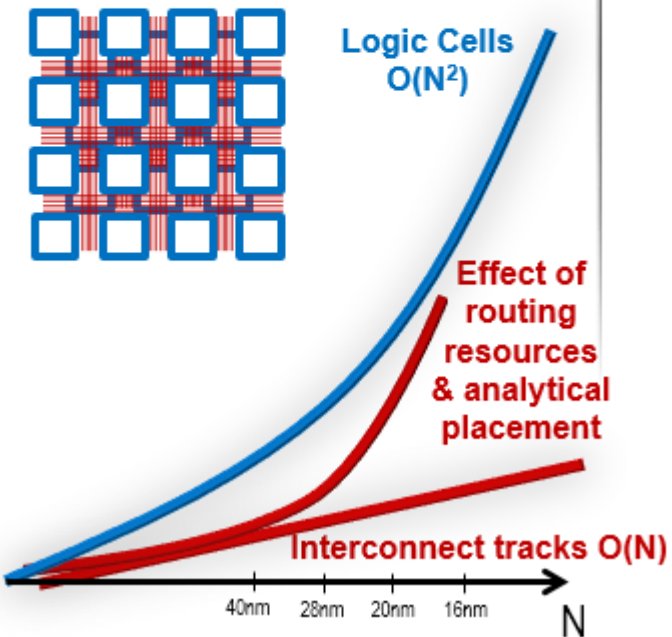


UltraScale Re-Architects the Core

Highest Utilization at Maximum Performance

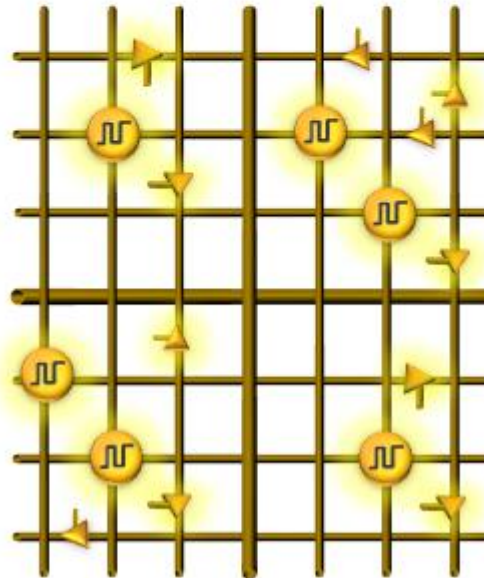
Next Generation Routing

- Re-designed routing architecture
- 2X routing, agile switching
- Co-Optimized with Vivado



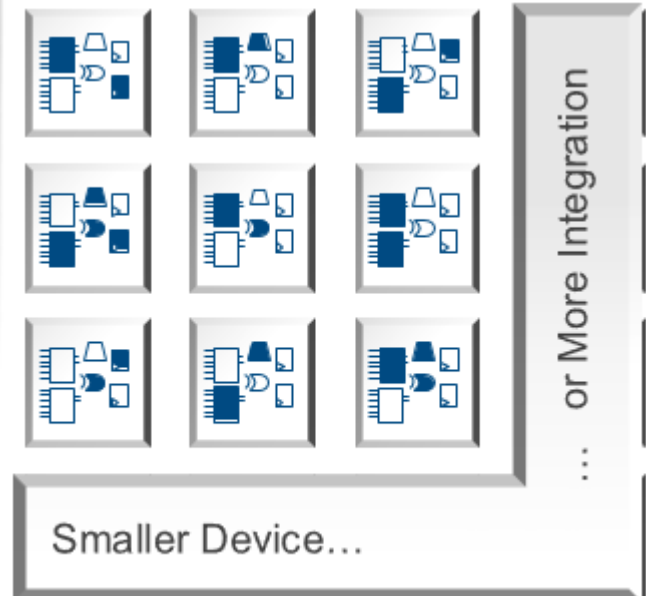
ASIC-Like Clocking

- Regional, segmented structure
- Flexible clock placement
- Scales w/density to balance skew

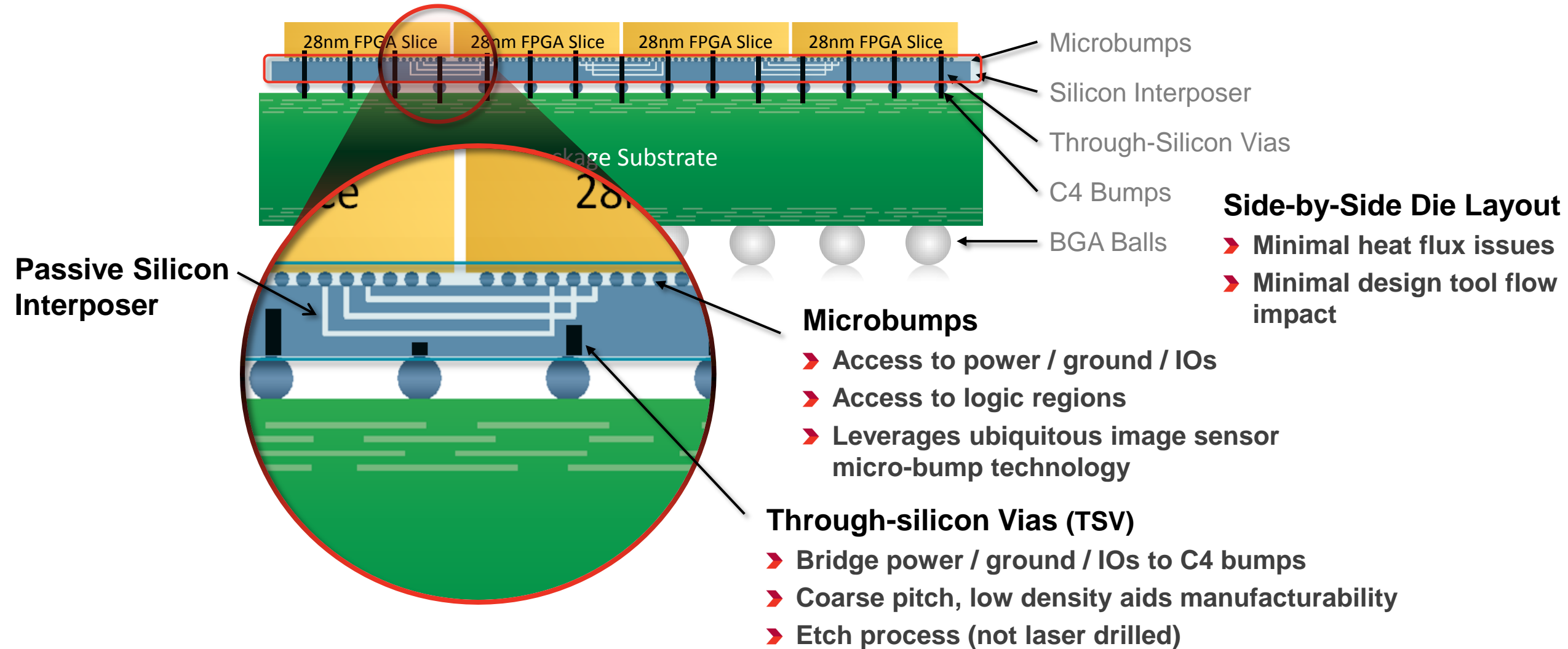


System Logic Cells

- Higher utilization enabled by routing
- Shorter net delays for performance
- Less wire switching for lower power

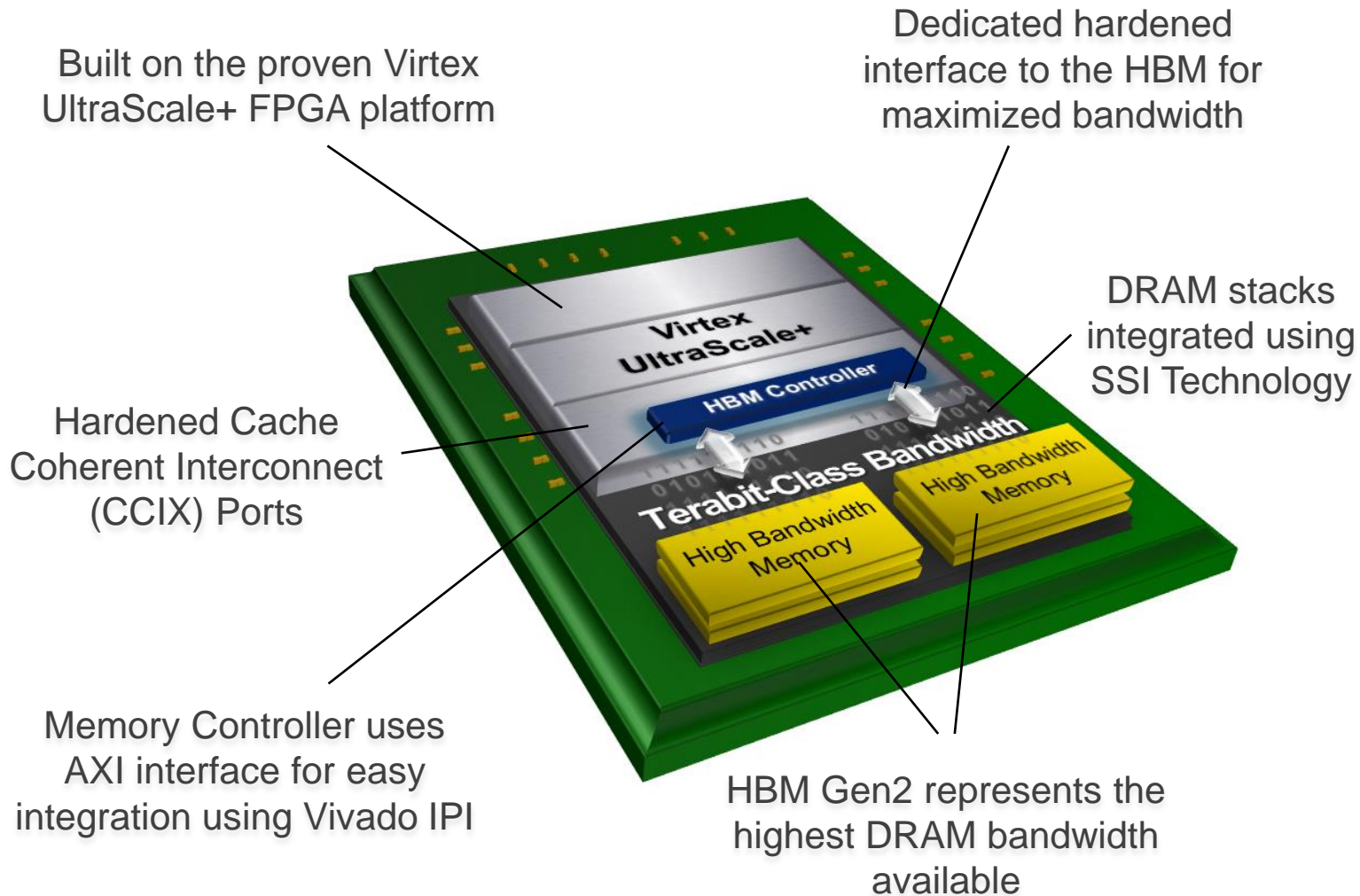


SSI Harnesses Proven Technology in a Unique Way

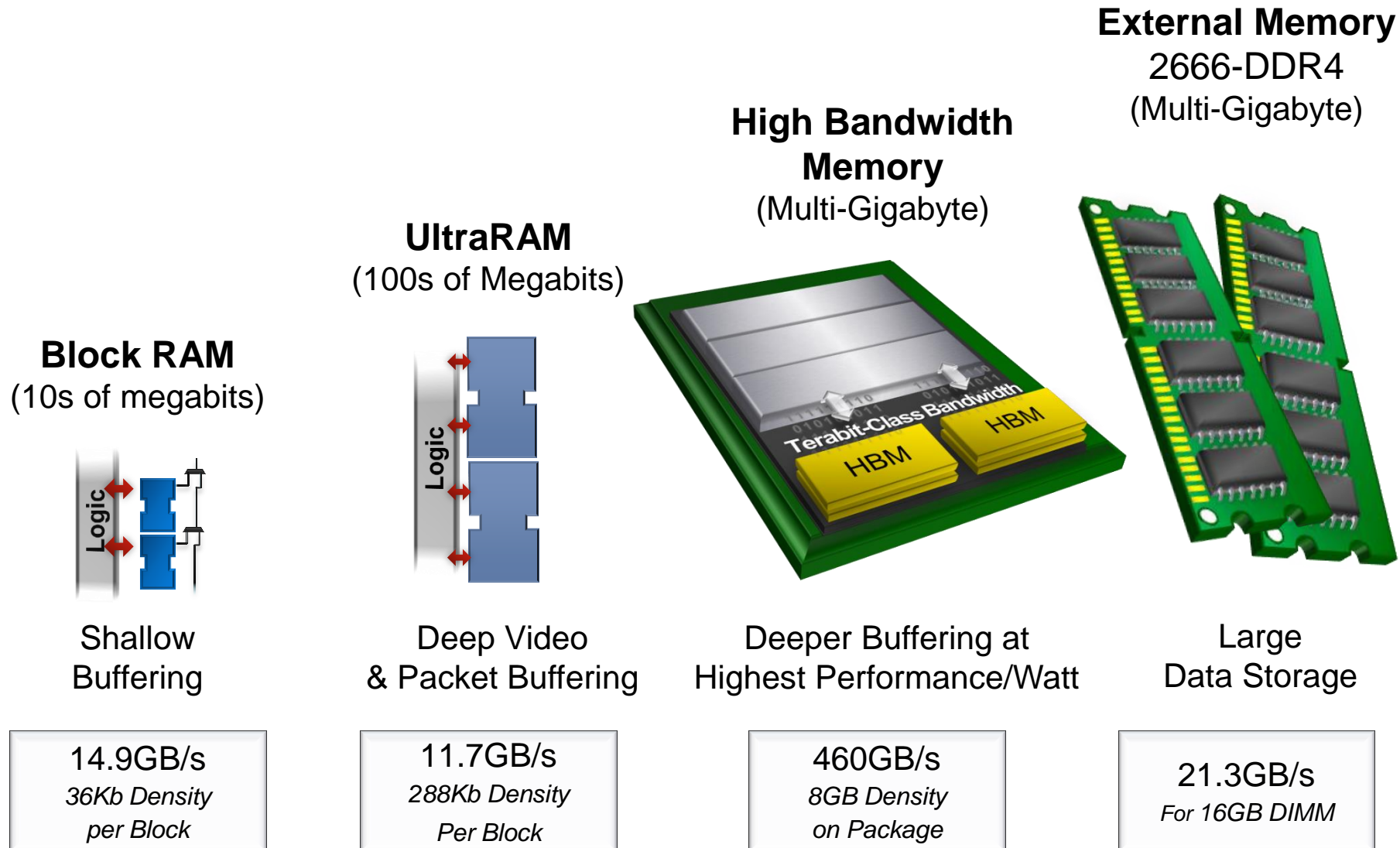


Introducing Virtex UltraScale+ HBM Devices

20X more bandwidth than a DDR4 DIMM

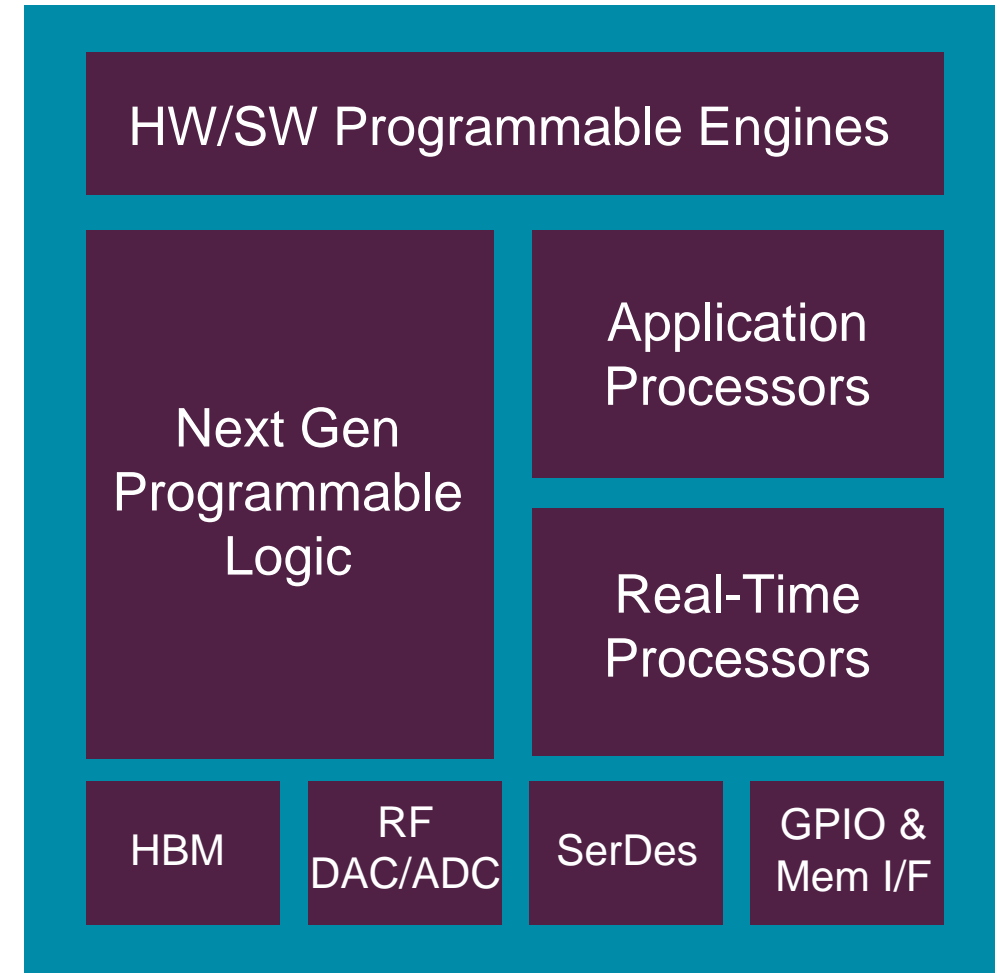


Covering the Full Spectrum of Memory Solutions



Next generation Adaptive Compute Acceleration Platform



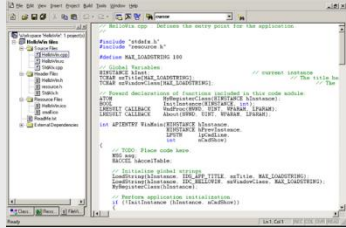
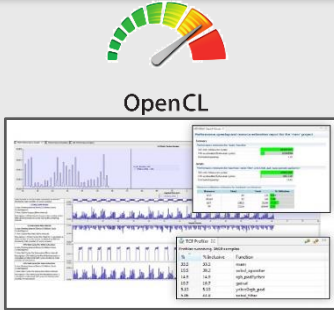
- New Device Category for Adaptive Workload-Specific Acceleration
- HW/SW programmable engines
- IP subsystems and a network-on-chip
- Highly integrated programmable I/O




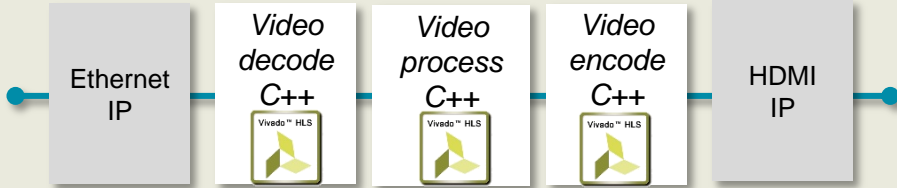
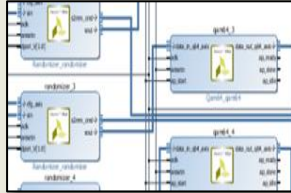
Software tools and libraries

Up-leveling the Programming Model


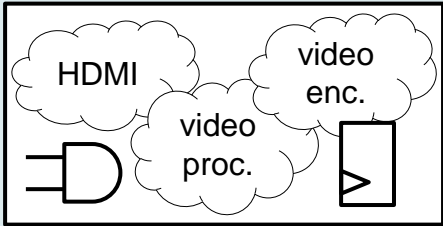
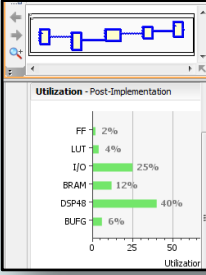
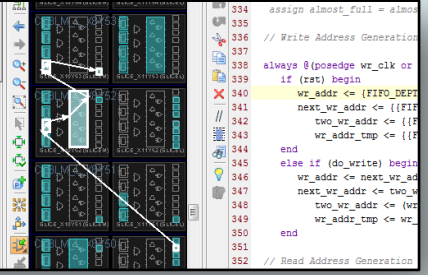
Development
Productivity

SW Programmability: SDAccel, SDSoC

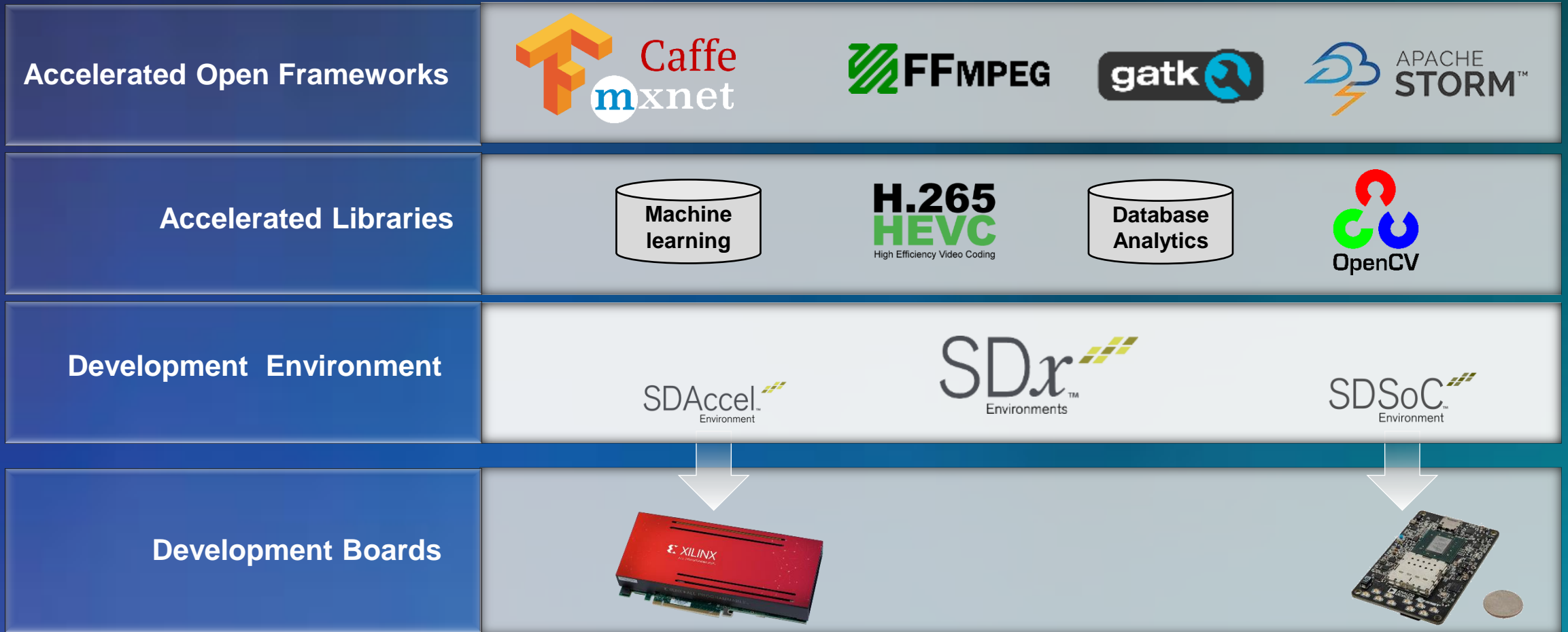




15x productivity: HLS, SysGen, Model Composer, IPI, SDK

Traditional HW design: Vivado, HDL

Development Stacks for SoC & FPGA

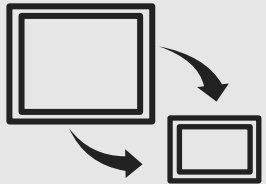


Advantages of Xilinx Devices



Reduced System Power Costs

- Highly efficient compute across range of workloads/applications
- Single device for range of processing & interfacing needs



Reduce System Hardware Costs

- FPGAs massive compute replaces CPU
- Future proofed hardware thanks to massive flexibility



Platform solutions

- Massive flexibility allows processing need for range of applications be met
- Massive IO flexibility ensures FPGA/SoC can hook into system easily
- Large range of devices to fit into system power envelop

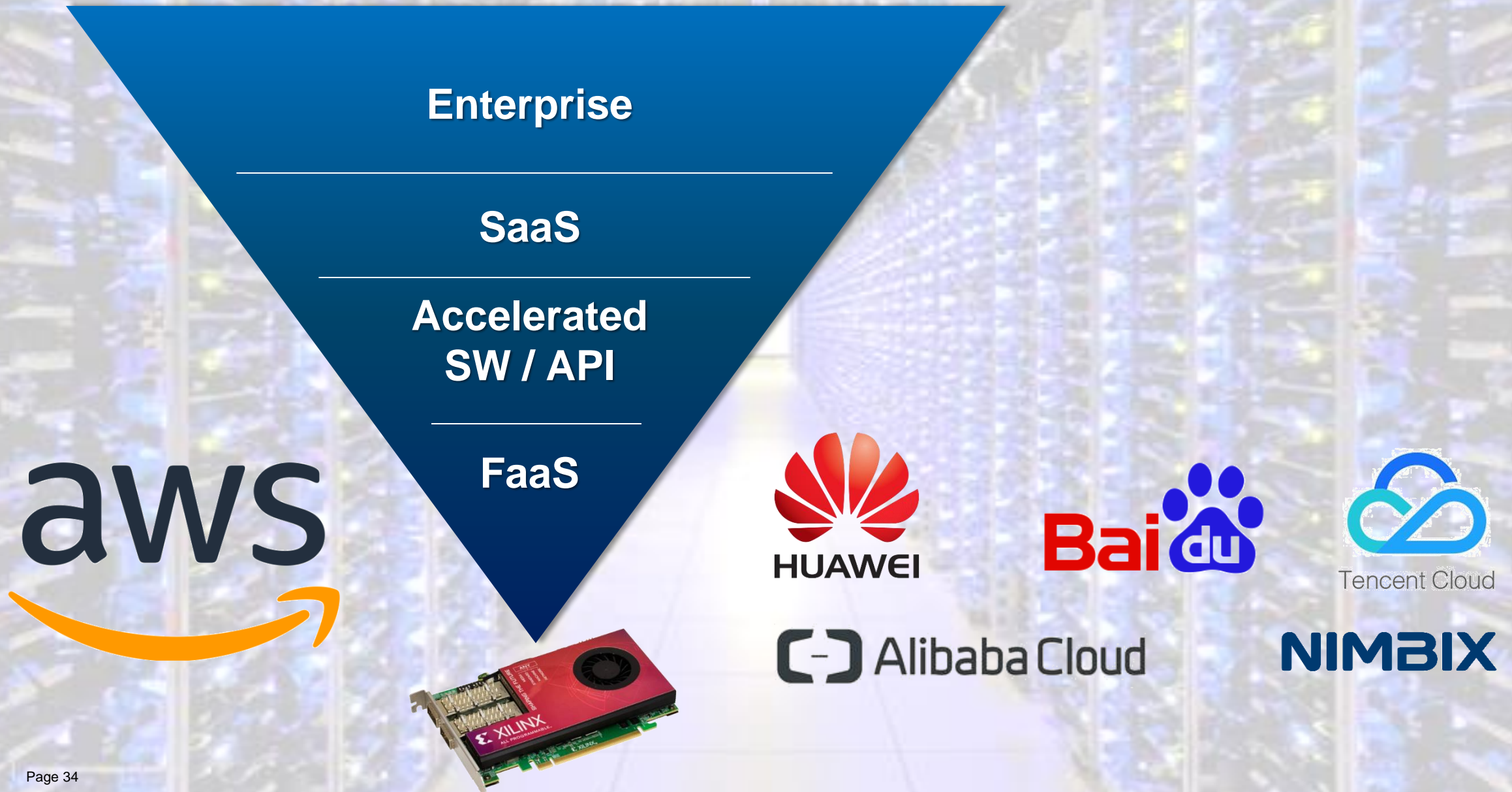


Ultimate Flexibility/Future Proofed HW

- Xilinx's flexibility & ability to handle diverse processing requirement
 - GPUs need high data locality, massive parallelism & specific data type + limited IO support
- Tomorrows algorithms will run efficiently on Xilinx devices

FPGAs in the cloud

FPGA as a Service Expanding Worldwide



Amazon F1 Instances



- EC2 Instance Type with up to 8 VU9P (16nmFF+) FPGA
- Intel Xeon Processors with up to 16 cores
- Connected through PCIe gen3x16
- 4 local DDR4 (12 GBps) per FPGA

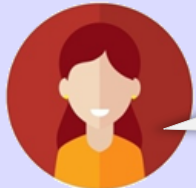
Model	#FPGA	Mem	SSD Storage	FPGA DDR4	Price / hour
f1.2xlarge	1	122 GB	470 GB	4x16 GB	\$1.65
f1.16xlarge	8	976 GB	8 x 470 GB	8 x 4x16 GB	\$13.20

F1 Users



end user

"I wasn't aware the service I am using involved F1 and FPGAs."



F1 developer #1

*"I need to **accelerate an application**. I don't know **RTL and hardware design**."*



F1 developer #2

*"I want to **create or reuse RTL** kernels while using **standard APIs** whenever possible."*



F1 developer #3

*"I want to create or reuse my **RTL designs** while designing HW and SW middleware."*



Tools

- User's front-end application leverages F1 transparently

- SDAccel

- ✓ Host: Xilinx OpenCL runtime
- ✓ Kernel: C, C++ or OpenCL

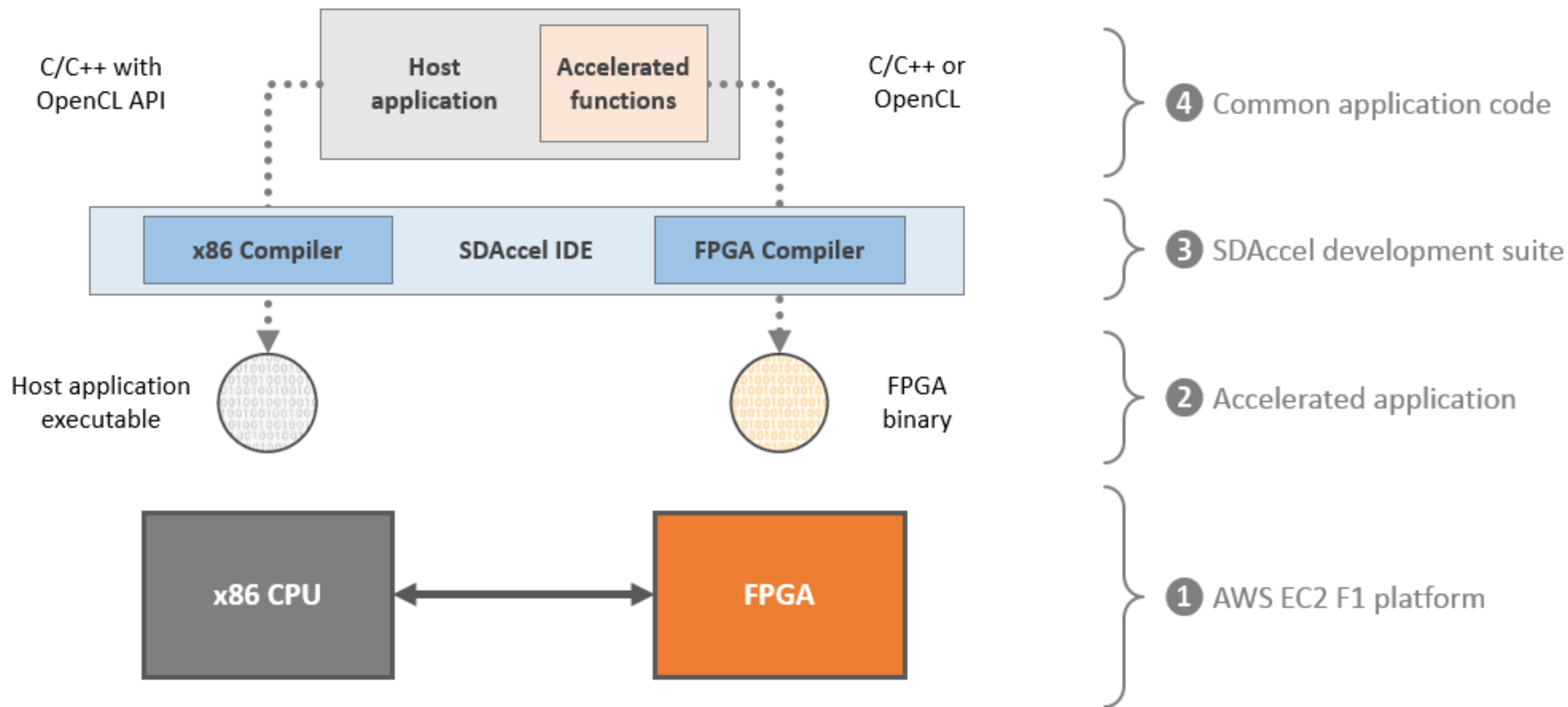
- SDAccel

- ✓ Host: Xilinx OpenCL runtime
- ✓ Kernel: RTL leveraging Vivado

- AWS HDK/SDK

- ✓ Host: Custom API
- ✓ Kernel: RTL or HLx

AWS EC2 F1 Instance SDAccel Flow



Xilinx tools in the cloud



FPGA Developer AMI

★★★★★ (3) | 1.3.3 | Sold by [Amazon Web Services](#)

\$0.00/hr for software + AWS usage fees

Linux/Unix, CentOS 7.3 | 64-bit Amazon Machine Image (AMI) | Updated: 9/24/17

The FPGA (field programmable gate array) AMI is a supported and maintained CentOS Linux image pre development tools and ...

[More info](#)

[Select](#)

Free tier eligible

➤ FPGA developer AMI

- Includes all Xilinx tools required for F1 development
- Run on standard AWS compute instances
- Spin up as many instances as you need on-demand
 - No Xilinx software/license maintenance
 - Reduces IT infrastructure requirements

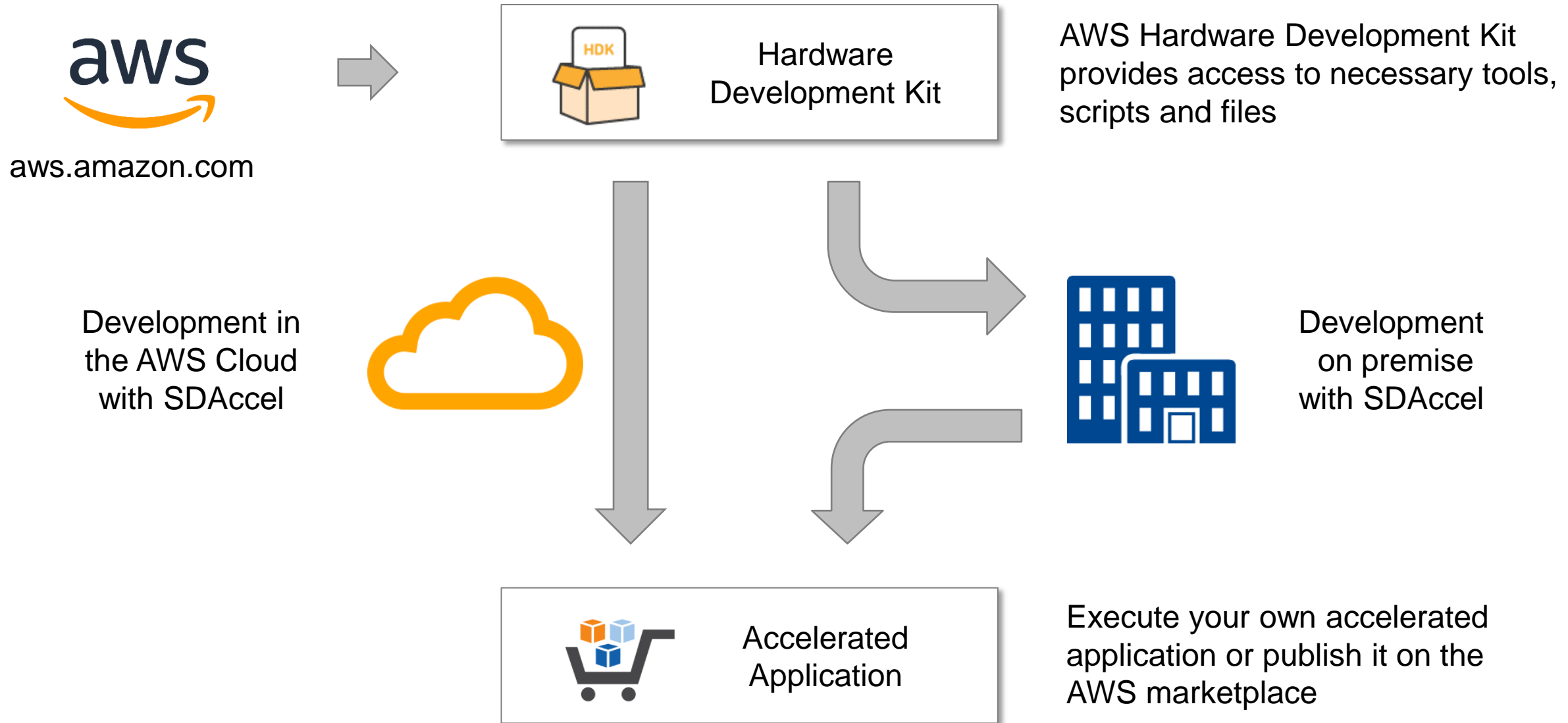
Filter by: Compute optimized Current generation [Show/Hide Columns](#)

Currently selected: c4.8xlarge (132 ECUs, 36 vCPUs, 2.9 GHz, Intel Xeon E5-2666v3, 60 GiB memory, EBS only)

Note: The vendor recommends using a **c4.4xlarge** instance (or larger) for the best experience with this product.

	Family	Type	vCPUs	Memory (GiB)
<input type="checkbox"/>	Compute optimized	c5.large	2	4
<input type="checkbox"/>	Compute optimized	c5.xlarge	4	8
<input type="checkbox"/>	Compute optimized	c4.xlarge	4	7.5
<input type="checkbox"/>	Compute optimized	c4.2xlarge	8	15
<input type="checkbox"/>	Compute optimized	c4.4xlarge	16	30
<input checked="" type="checkbox"/>	Compute optimized	c4.8xlarge	36	60

Amazon F1 Development Flow



Benefits of the AWS F1 Cloud Compute Platform

- Accelerated computation
- Makes leading-edge FPGA acceleration available to a large community of developers, and to millions of potential users
- Provides dedicated and large amounts of leading-edge FPGA logic with elasticity to scale to multiple instances
- Simplifies the development process by providing cloud-based tools for FPGA development
- Ideal platform for collaborative research and development – build a prototype and share instantly with partners

Workloads



Video

10x



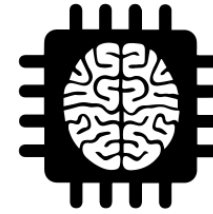
Data analytics

90x



Genomics

100x



Machine
Learning

40x



& more



Reconfigure.io



Saving Babies at Cloud Scale

edico  genome

20min genome analysis (100x faster)

Rady Children's Hospital-San Diego



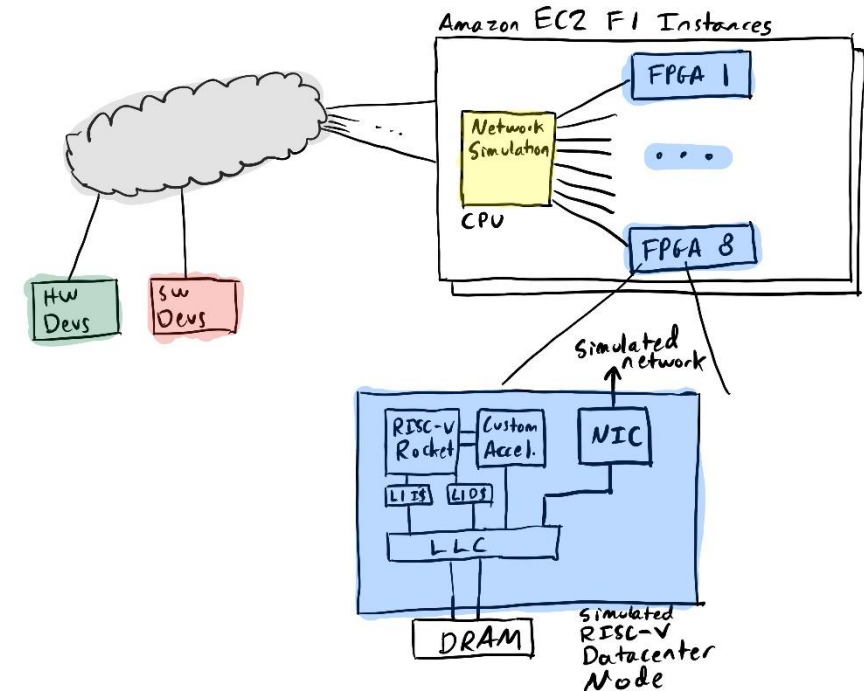
Fastest Analysis of 1000 Whole Human Genomes
in 25min on AWS



FireSim - Cycle-accurate, FPGA-accelerated data center simulation project based on RISC-V

➤ Uses public-cloud F1

- No upfront cost to purchase and deploy FPGA hardware
- Distribute pre-built images
 - Easy to reproduce experiments*
 - Automates FPGA simulation
- Scale out experiments by spinning up additional EC2 instances
- “Saves hundreds of thousands of dollars on large FPGA clusters”



*Try for yourself:

FireSim

FireSim Demo v1.0

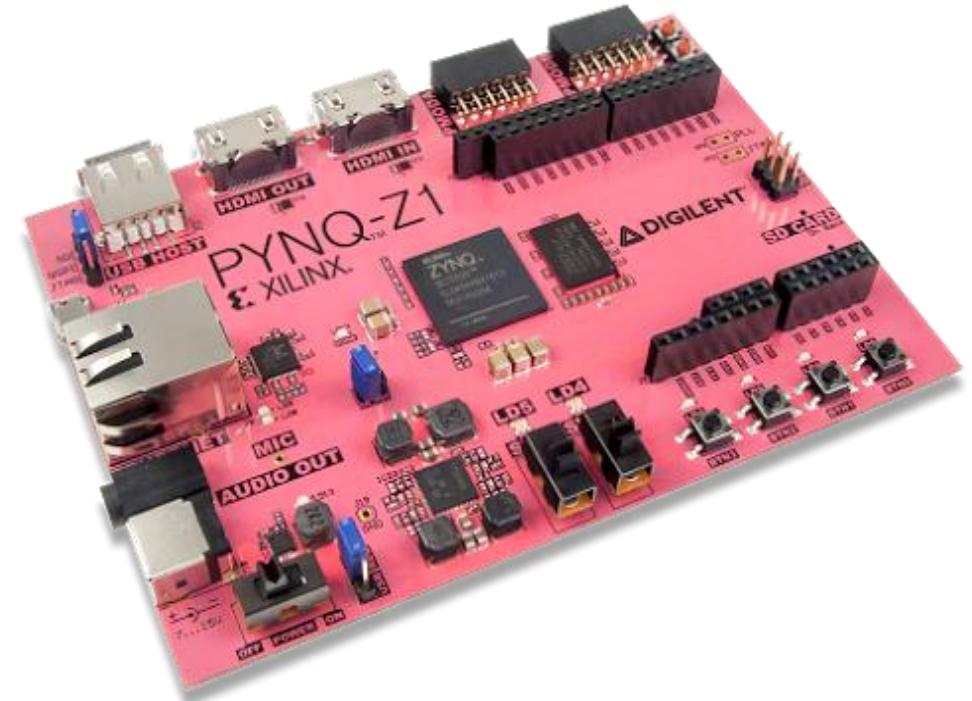
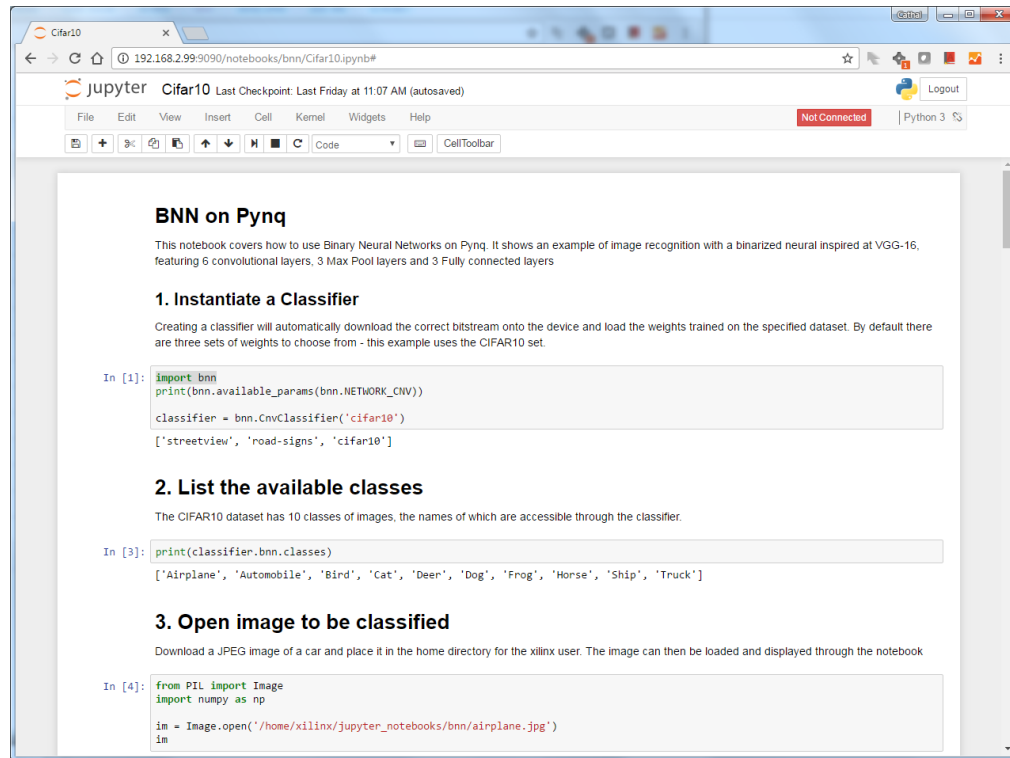
Sold by: [Berkeley Architecture Research](#) Latest Version: 1.0

This image includes an AMI and AFI to demo FireSim, a fast, cycle-accurate FPGA-accelerated hardware simulation tool. This release can simulate a single-node or eight-node cluster of

From cloud to edge and back



"Python Productivity for Zynq"



Linux

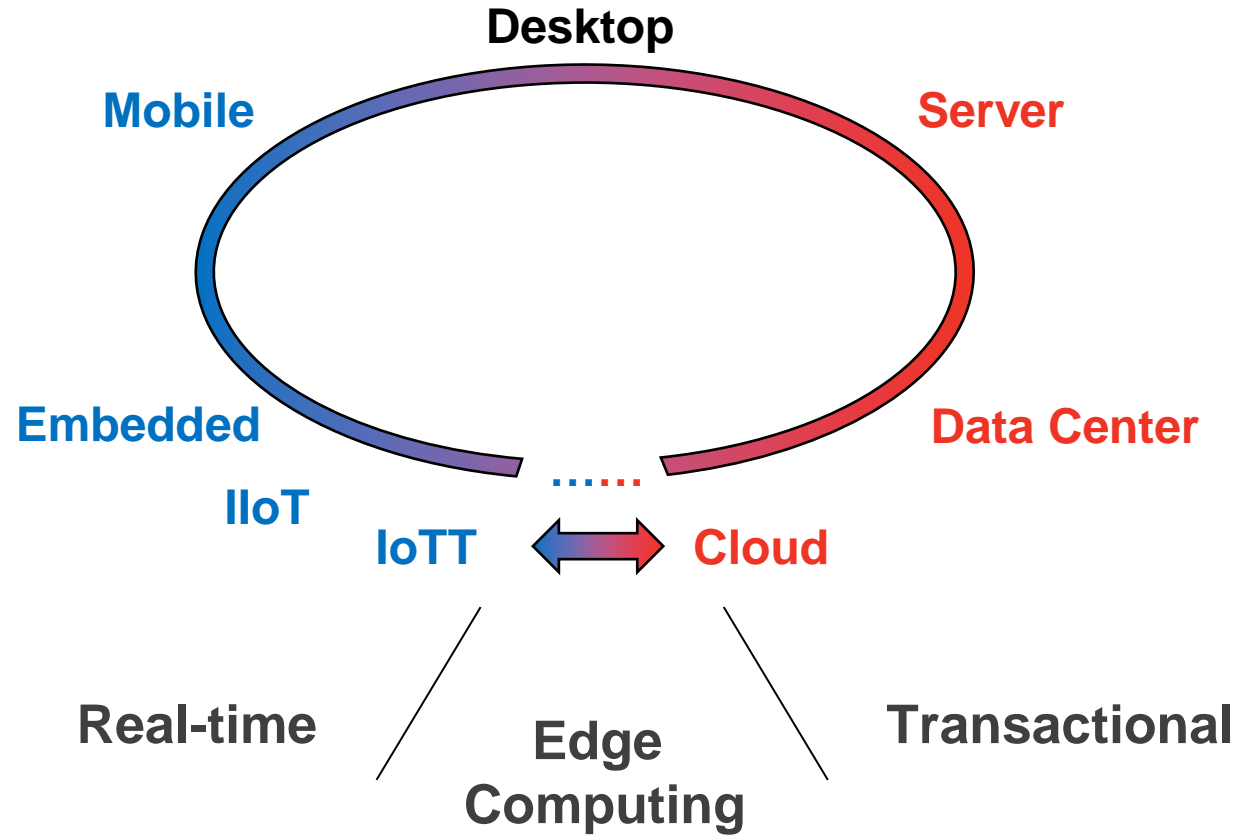
Computing Landscape: ... A Linear Spectrum from IOT to Cloud

IoTT .. IIoT .. Embedded .. Mobile .. Desktop .. Server .. Data Center .. Cloud

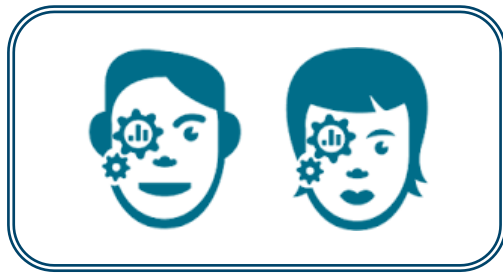


- IIoT: Industrial Internet of Things
- IoTT: Internet of Tiny Things (aka motes, ultra low-power/energy)

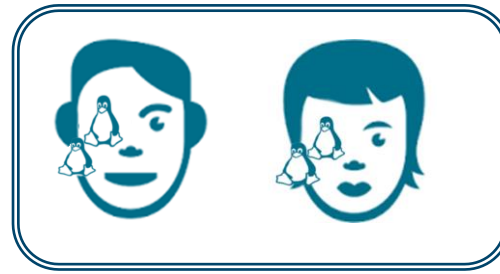
Computing Landscape: ... From Linear Spectrum to Continuum



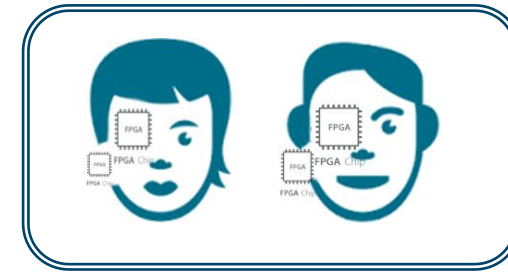
PYNQ enables hardware, software and analytics



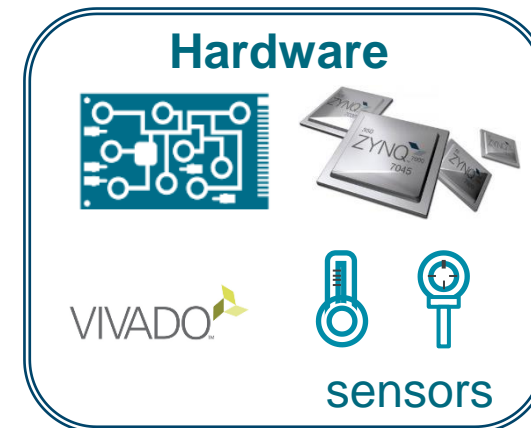
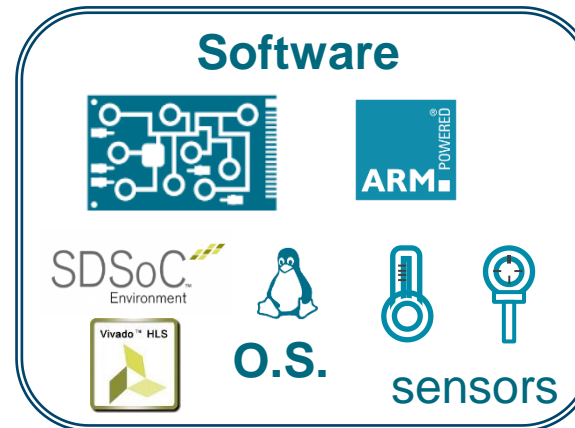
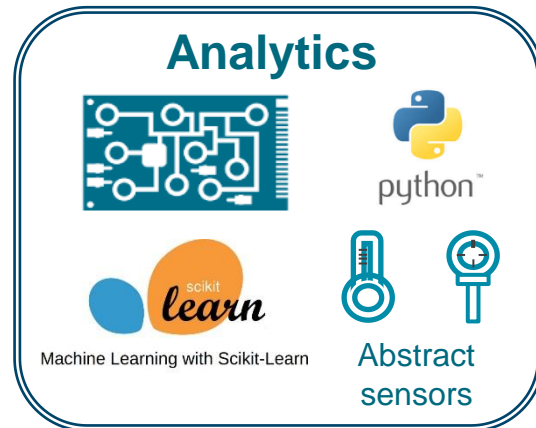
Data Scientists



Embedded Engineers

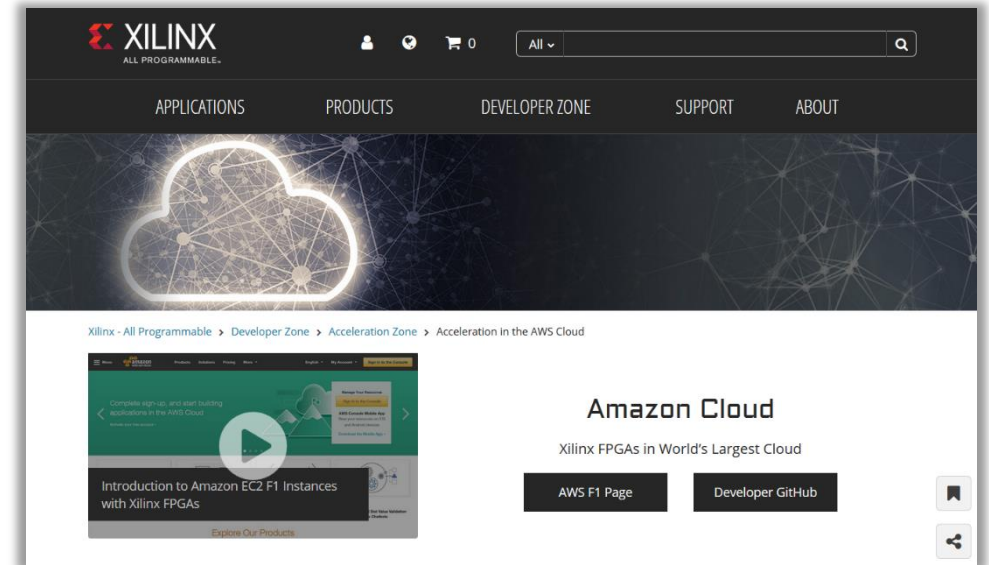


Hardware Engineers



Summary

- Delivering more than Moore
 - Ultrascale+, Everest
- Adaptable computing for AI
- Cloud solutions for big data problems
 - AWS EC2 F1
- From cloud to edge and back with PYNQ



<https://www.xilinx.com/products/design-tools/acceleration-zone/aws.html>

PYNQ™

