Enabling Technology for the Cloud and AI – One Size Fits All?

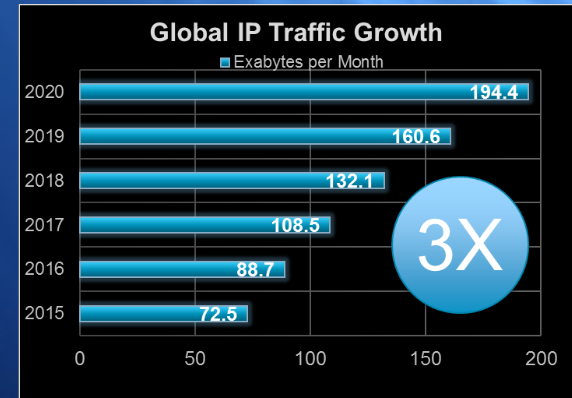
eSilicon®

Collaborate. Differentiate. Win.

Tim Horel

DIRECTOR, FIELD APPLICATIONS

The Growing Cloud – Global IP Traffic Growth



Reference: Cisco VNI Global IP Traffic Forecast, 2015 - 2020

40B+ devices with intelligence will ship in 2020

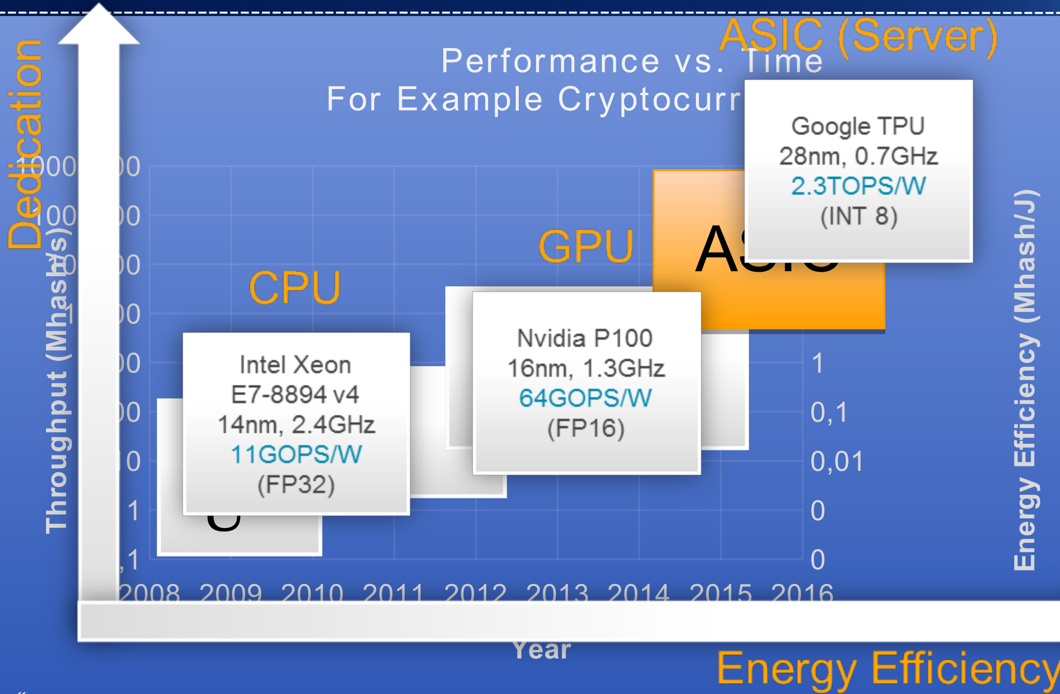
1B+ additional consumers online, 30B+ connected devices

~200 Exabytes of data traffic per month

© 2018 eSilicon Corporation. All rights reserved

ASIC Offers Performance Improvement

eSilicon



"The Future of Machine Learning Hardware", Phillip Jama, Sept. 2016,
https://hackernoon.com/the-future-of-machine-learning-hardware-c872a0448be8

These numbers are based on the performance factors (such as throughput, efficiency) observed through the cryptocurrency-mining evolution.

- *Non-specialized hardware comparison* https://en.bitcoin.it/wiki/Non-specialized_hardware_comparison
- *Mining hardware comparison* https://en.bitcoin.it/wiki/Mining_hardware_comparison

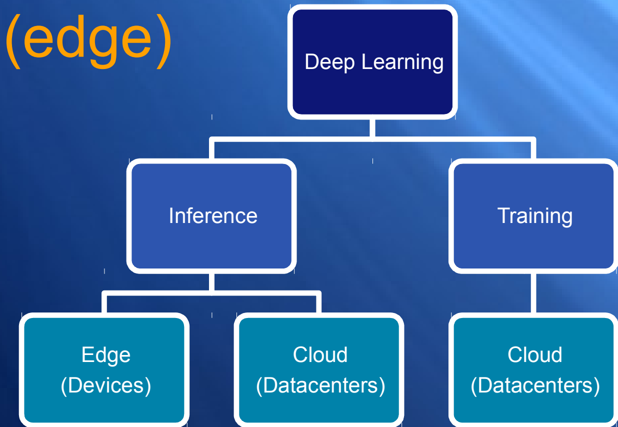
- Performance measurement:
  - Computational capacity (or throughput)
  - Energy-efficiency (computations per Joule)
  - Cost-efficiency (throughput per dollar)
- Computational units:
  - Single-core CPU:1
  - Multi-core CPU: 10
  - GPU: 100
  - FPGA: 1,000
  - ASIC: 10,000 to 1,000,000

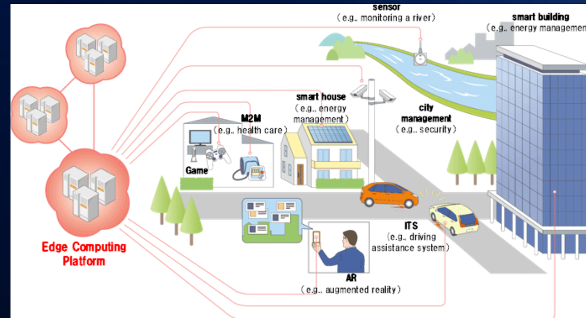
© 2018 eSilicon Corporation. All rights reserved

Analytics in the Cloud with Artificial Intelligence

Deep learning is deployed today in all major datacenters (cloud) and in many devices (edge)

- Flood of data – images, video, text, transactions, speech
- Hardware acceleration enables and scales deep learning computation



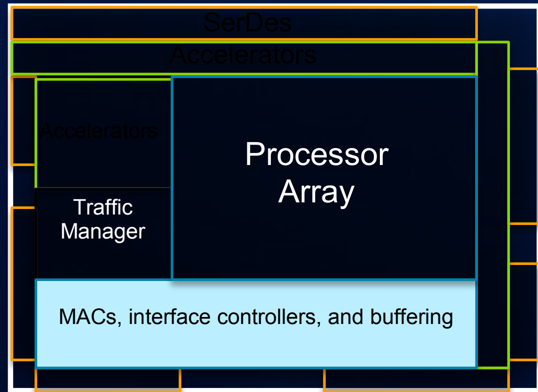


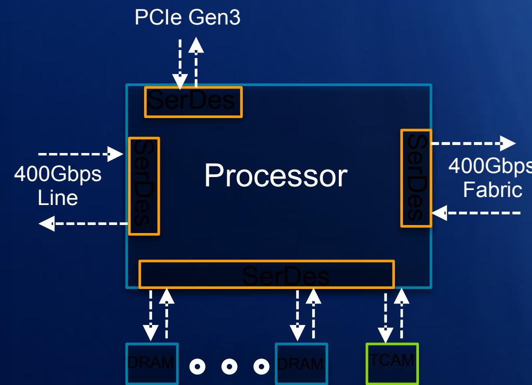
Source: NTT Group

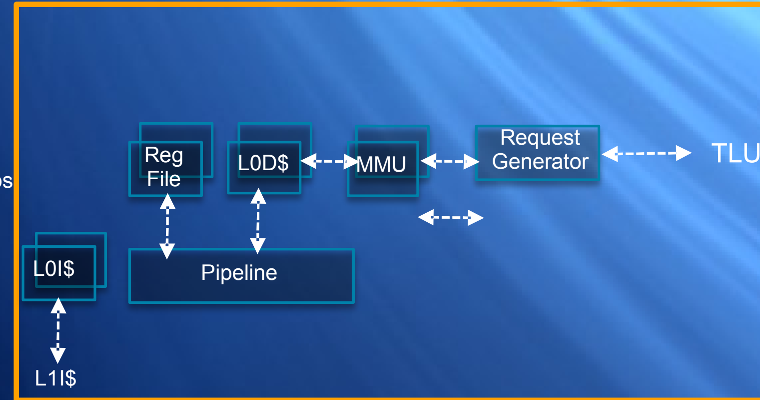
Source: Datacenter Frontier

© 2018 eSilicon Corporation. All rights reserved

Example - 400Gbps Multi-Core Network Processor







- 9.2 billion transistors; 643mm² die
- 672 total processors: 42 processor clusters
  - Link out to a four-way set associative L2 instruction
  - On-chip interconnect that links the clusters and caches to each
- 276 SerDes
- Interconnect runs at 1 GHz
- 9 Tb/sec of aggregate bandwidth

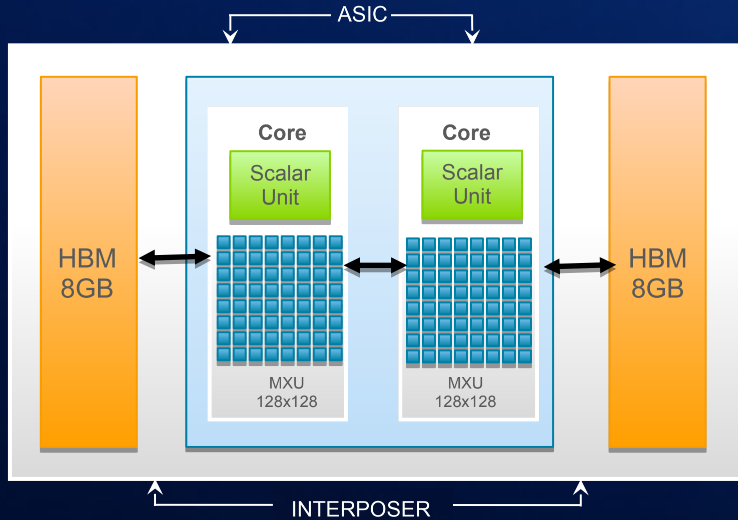
- 353 Mb of SRAM per core
  - L0 data and instruction cache for each thread
  - L1 instruction and data caches (for a cluster of sixteen cores)
- External DRAM for large data structures and packet buffering
- External TCAM for large data structures
- Integrated Ethernet MACs from 10GE to 100GE

Reference: «A 400Gbps Multi-Core Network Processor», James Markevitch, Srinivasa Malladi Cisco Systems, August 22, 2017

© 2018 eSilicon Corporation. All rights reserved

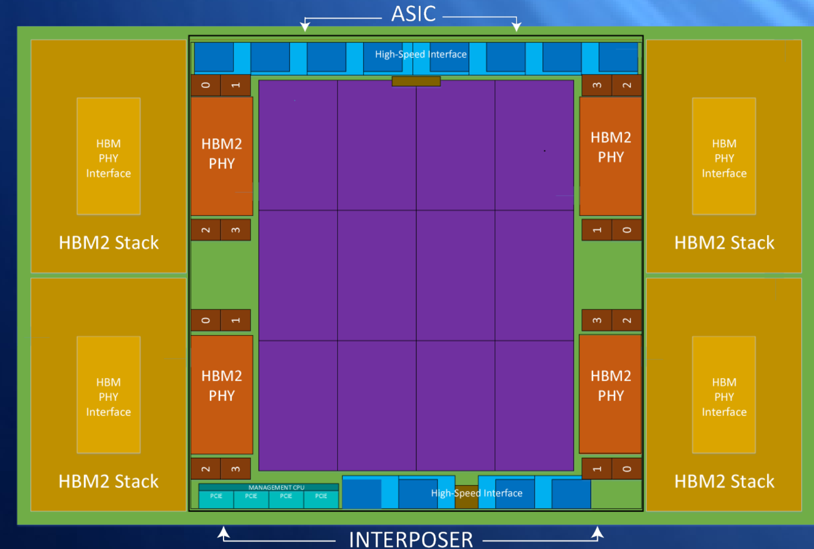
Example AI Processors In The Market Today

eSilicon



- MXU: 32b float accumulation but reduced precision for multipliers
- 45 TFLOPS
- 16GB HBM
- 600GB/s memory bandwidth
- Scalar unit: 32b float

"Case Study on the Google TPU and GDDR5 from Hot Chips 29", Patrick Kennedy, Aug. 2017



- Tensor-based architecture
- Flexpoint® – high parallelism, low power
- 4X HBM2 memory stacks

"https://www.anandtech.com/show/11942/intel-shipping-nervana-neural-network-processor-first-silicon-before-year-end", Nate Oh, October 2017

© 2018 eSilicon Corporation. All rights reserved

Similar Yet Different

eSilicon

Networking

Mature

Well established architectures with products that have migrated for multiple technology generations

Immature

General purpose chips offer a costly rapid-to-deployment option

Evolving algorithms make ASICs harder to define

ASICs designed today will serve a more narrow application tomorrow

AI

© 2018 eSilicon Corporation. All rights reserved

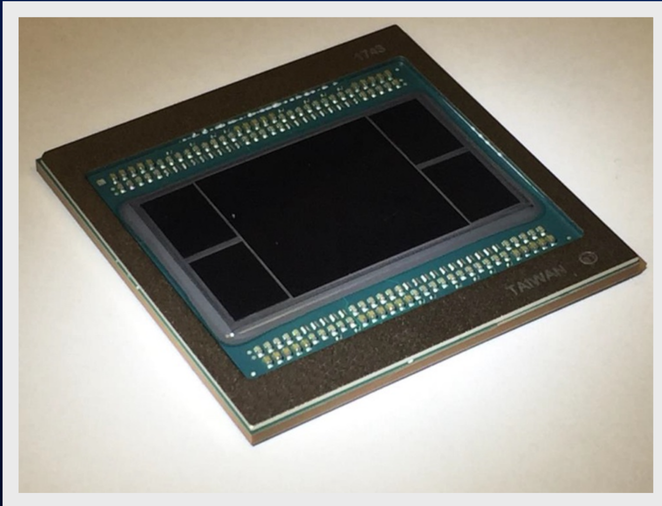
Networking ASIC with High-Bandwidth Memory

3.6 Tb/sec

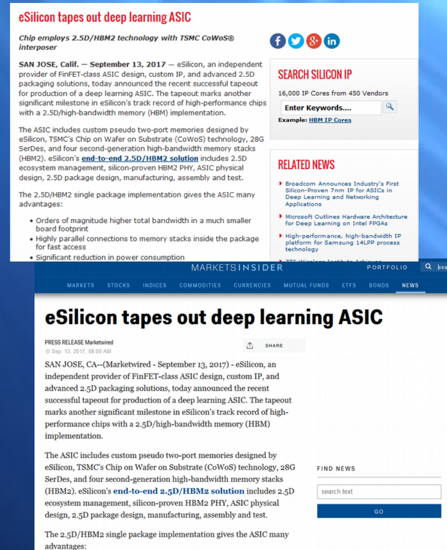
- FinFET technology
- 150M gates
- 900Mb embedded SRAM
- One 4-high HBM2 stack
- 60 lanes multi-protocol SerDes, 1 to 28Gb
- HBM2 PHY
- Memory compilers
- Specialty I/O
- Custom memory instances optimized for power/performance/area
- TCAM (supports 1.2G search/second [GSPS])

© 2018 eSilicon Corporation. All rights reserved

Deep Learning ASIC with High Bandwidth Memory and Custom Memory



- 8 Terabits/s of memory access speed
- ~400Mb embedded SRAM
- 4 HBM2 stack - 1TB/s HBM bandwidth
- 2.5D Packaging
- 28Gb SerDes
- HBM Gen2 PHY
- Memory compilers
  - 2PRF, DP SRAM, Fast Cache
- Deep Learning specialty SRAM
  - Custom Pseudo 2-Port SRAM



eSilicon tapes out deep learning ASIC

Chip employs 2.5D/HBM2 technology with TSMC CoWoS® interposer

SAN JOSE, Calif. — September 13, 2017 — eSilicon, an independent provider of FinFET-class ASIC design, custom IP, and advanced 2.5D packaging solutions, today announced the recent successful tapeout for production of a deep learning ASIC. The tapeout marks another significant milestone in eSilicon's track record of high-performance chips with a 2.5D/high-bandwidth memory (HBM) implementation.

The ASIC includes custom pseudo two-port memories designed by eSilicon, TSMC's Chip on Wafer on Substrate (CoWoS) technology, 28G SerDes, and four second-generation high-bandwidth memory stacks (HBM2). eSilicon's end-to-end 2.5D/HBM2 solution includes 2.5D ecosystem management, silicon-proven HBM2 PHY, ASIC physical design, 2.5D package design, manufacturing, assembly and test.

The 2.5D/HBM2 single package implementation gives the ASIC many advantages:

- Orders of magnitude higher total bandwidth in a much smaller board footprint
- Highly parallel connections to memory stacks inside the package for fast access
- Significant reduction in power consumption

SEARCH SILICON IP

16,000 IP Cores from 450 Vendors

Enter Keywords....

Example: HBM IP Cores

RELATED NEWS

- Broadcom Announces Industry's First Silicon-Proven 7nm IP for ASICs in Deep Learning and Networking Applications
- Microsoft Outlines Hardware Architecture for Deep Learning on Intel FPGAs
- High-performance, high-bandwidth IP platform for Samsung 14LPP process technology

MARKETSINSIDER PORTFOLIO

MARKETS STOCKS INDICES COMMODITIES CURRENCIES MUTUAL FUNDS ETFS BONDS NEWS

eSilicon tapes out deep learning ASIC

PRESS RELEASE Marketwired
Sep. 13, 2017, 08:00 AM

SHARE

SAN JOSE, CA--(Marketwired - September 13, 2017) - eSilicon, an independent provider of FinFET-class ASIC design, custom IP, and advanced 2.5D packaging solutions, today announced the recent successful tapeout for production of a deep learning ASIC. The tapeout marks another significant milestone in eSilicon's track record of high-performance chips with a 2.5D/high-bandwidth memory (HBM) implementation.

The ASIC includes custom pseudo two-port memories designed by eSilicon, TSMC's Chip on Wafer on Substrate (CoWoS) technology, 28G SerDes, and four second-generation high-bandwidth memory stacks (HBM2). eSilicon's end-to-end 2.5D/HBM2 solution includes 2.5D ecosystem management, silicon-proven HBM2 PHY, ASIC physical design, 2.5D package design, manufacturing, assembly and test.

The 2.5D/HBM2 single package implementation gives the ASIC many advantages:

FIND NEWS

search text

GO

News Feed Item

eSilicon tapes out deep learning ASIC

Chip employs 2.5D/HBM2 technology with TSMC CoWoS(R) interposer

BY MARKETWIRED . ARTICLE RATING:

SEPTEMBER 13, 2017 08:00 AM EDT

RELATED PRINT EMAIL FEEDBACK ADD THIS BLOG THIS

MARKET WIRED eSilicon

SAN JOSE, CA--(Marketwired - September 13, 2017) - eSilicon, an independent provider of FinFET-class ASIC design, custom IP, and advanced 2.5D packaging solutions, today announced the recent successful tapeout for production of a deep learning ASIC. The tapeout marks another significant milestone in eSilicon's track record of high-performance chips with a 2.5D/high-bandwidth memory (HBM) implementation.

The ASIC includes custom pseudo two-port memories designed by eSilicon, TSMC's Chip on Wafer on Substrate (CoWoS) technology, 28G SerDes, and four second-generation high-bandwidth memory stacks (HBM2). eSilicon's end-to-end 2.5D/HBM2 solution includes 2.5D ecosystem management, silicon-proven HBM2 PHY, ASIC physical design, 2.5D package design, manufacturing, assembly and test.

The 2.5D/HBM2 single package implementation gives the ASIC many advantages:

Comments

New Release of Quest Toad for Oracle Offers Enhancements to Reduce Risk

By Liz McMillan

yourfanat wrote: I am using another tool for Oracle developers - dbForge Studio for Oracle. This IDE has lots of usefull features, among them: oracle designer, code competion and formatter, query builder, debugger, profiler, erxport/import, reports and many others. The latest version supports Oracle 12C. More information here.

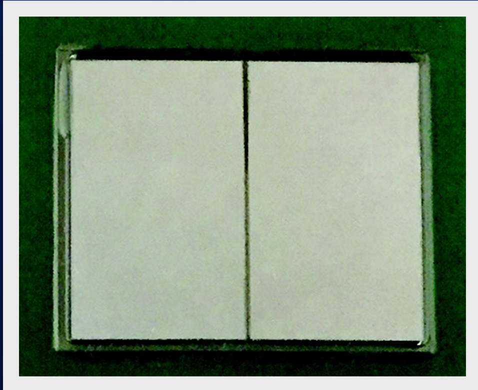
Oct. 22, 2013 02:57 AM EDT

read more & respond »

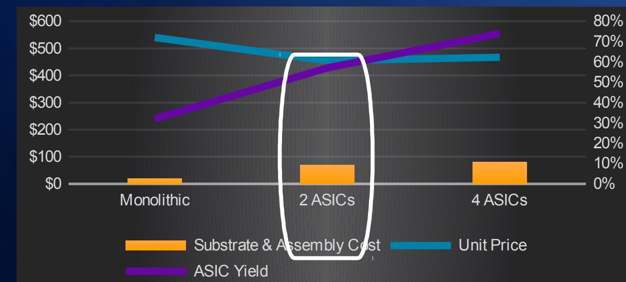
© 2018 eSilicon Corporation. All rights reserved

Economics of Billions of Transistors: Divide and Conquer

eSilicon



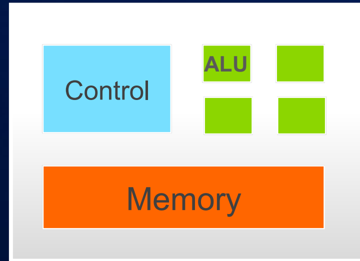
- FinFET technology
- Die-to-die interconnect on interposer running at >2.5Gbps per pin (~6000 interconnects between two die)
- Leverages HBM PHY technology, adopted for die-to-die
- Pre-requisite: Architect for yield optimization

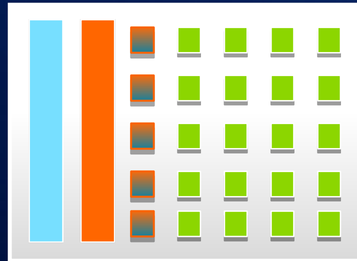


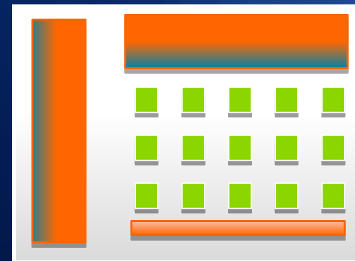
© 2018 eSilicon Corporation. All rights reserved

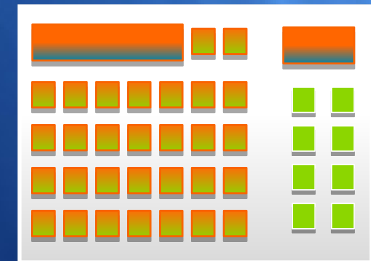
Processing Units – Rapid and Diverse Innovation Area

eSilicon









CPU
- Complex control logic
- Low compute density
- High programmability

→ Parallel Processing

GPU
- Parallel computation
- High compute density
- Floating point datatype

→ Dedicated Architecture

ASIC
- Matrix multiply arch.
- Dedicated memory arch.
- Fixed point datatype

→ Heterogeneous Architecture

DNPU
- Heterogeneous arch.
- Fixed computation pattern
- Dedicated datatype

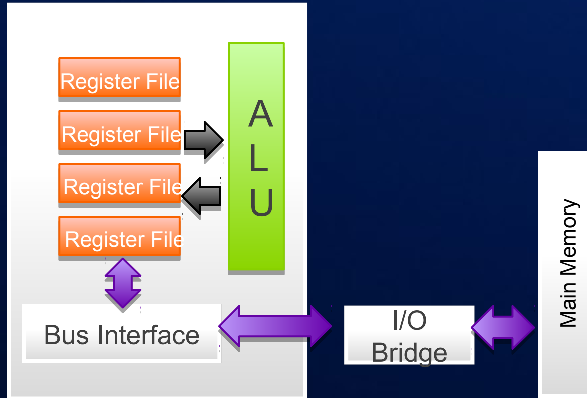
Reference: "DNPU: An Energy-Efficient Deep Neural Network Processor with On-Chip Stereo Matching", Dongjoo Shin, Jinmook Lee, Jinsu Lee, Juhyoung Lee, and Hoi-Jun Yoo, Semiconductor System Laboratory School of EE, KAIST

© 2018 eSilicon Corporation. All rights reserved

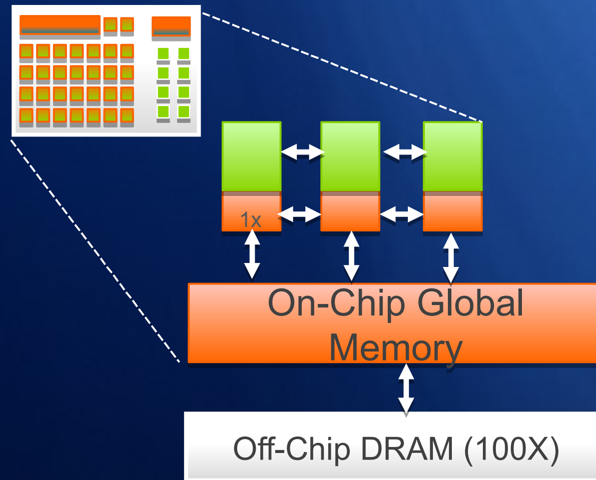
Latency v. Throughput at the Cost of Power

eSilicon

Traditional Compute Memory Subsystem



Today's Memory Subsystem



- Data movement has high energy cost
- Processing elements stalled waiting for data
- Processors optimize data flow to reduce energy cost
- *Off-chip memory is 100X less energy efficient than local memory*

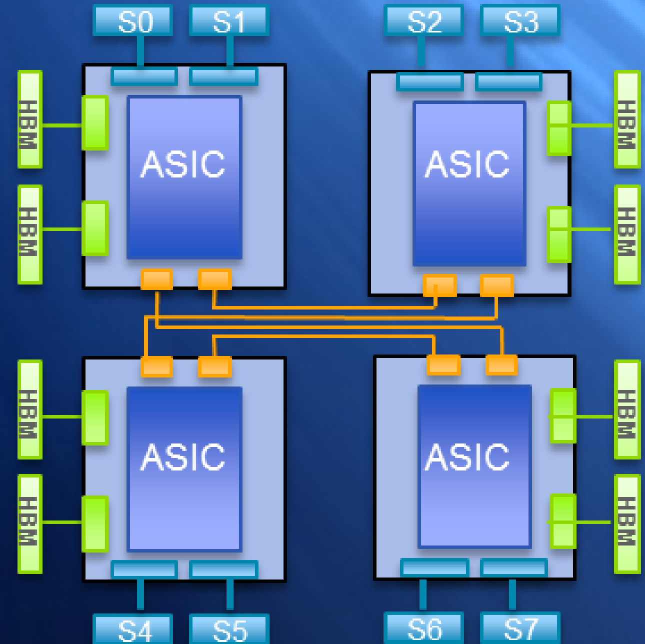
- More on-chip heterogeneous SRAM
- Traditional "one size fits all" embedded SRAM is no longer applicable for both Networking and AI
- Operations can be done in parallel, however memory access and power is key bottleneck!

© 2018 eSilicon Corporation. All rights reserved

"Hardware Trends and their Implications", University of Glasgow, School of Computing Science

Interconnect – Lacing the Fabric

- Chip-to-chip interconnect
  - Short reach multi-chip modules
    - Packaging of dies with dissimilar semiconductor processes
    - Packaging of smaller dies to increase yield
  - No standard, proprietary solutions
- High throughput data interfaces:
  - Processor or switch to wide, high bandwidth memory
  - Off-board SerDes - 56G/112G SerDes PHY for long-reach backplane requirements for the 400 GB Ethernet
  - Switch to switch links



© 2018 eSilicon Corporation. All rights reserved

Rapid Pace of Innovation

eSilicon

Quick Evolution

- Short time to product to hit tight market windows
- Architectural flexibility to allow for incremental refinements
- High degree of re-use to shorten design cycle
- Swappable modules and heterogeneous technologies
- “What if” exploration capabilities

Virtual Acceleration

Unique IP offering based on mathematical functions integrated with memory

Highly parallel, embedded interconnects

Very scalable solutions across multiple dies, chiplets, packages and technologies

© 2018 eSilicon Corporation. All rights reserved

Shorten Time To Tape-out for AI

eSilicon

Library of Mega Cells

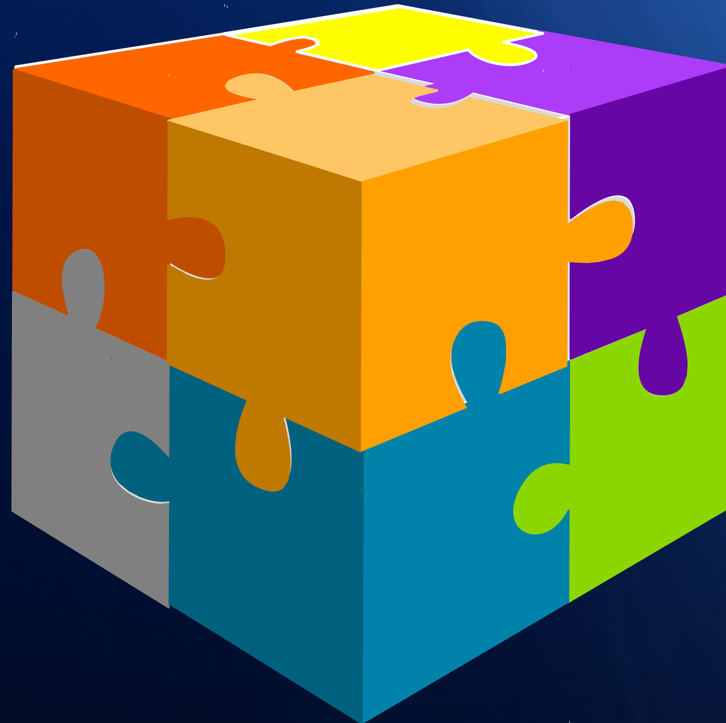
Configurable, hardened cells, referenced from RTL, specialized memory

On-Chip Interconnect scheme

Configurable NOC

High Speed Off-chip Interconnect

Configurable, hardened high-speed interconnect , 56/112G SerDes, HBI



Off-Chip Memory

HBM2 DRAM, PHY/controller hardened and pre-validated

2.5 D Packaging

Interposer design, signal & thermal integrity, package warpage

ASIC Toolbox & Recipe

Database and tools to analyze, assess and determine most optimal technology, logic and memory implementation

© 2018 eSilicon Corporation. All rights reserved

eSilicon®

Collaborate. Differentiate. Win.