Pivotal Memory Technologies Enabling New Generation of AI Workloads

Tien Shiah

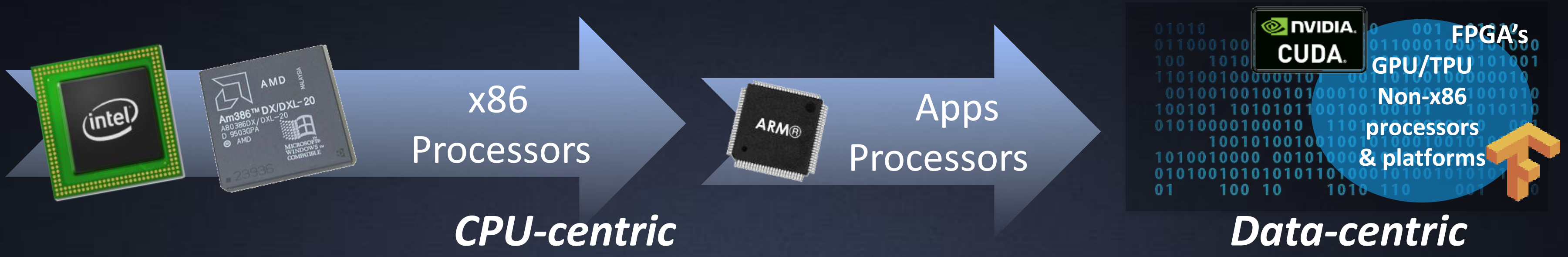
Memory Product Marketing
Samsung Semiconductor Inc.

Legal Disclaimer

This presentation is intended to provide information concerning the memory industry. We do our best to make sure that information presented is accurate and fully up-to-date. However, the presentation may be subject to technical inaccuracies, information that is not up-to-date or typographical errors. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of information provided on this presentation.

The information in this presentation or accompanying oral statements may include forward-looking statements. These forward-looking statements include all matters that are not historical facts, statements regarding the Samsung Electronics' intentions, beliefs or current expectations concerning, among other things, market prospects, growth, strategies, and the industry in which Samsung operates. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements contained in this presentation or in the accompanying oral statements. In addition, even if the information contained herein or the oral statements are shown to be accurate, those developments may not be indicative of developments in future periods.

Applications drive Changes in Architectures



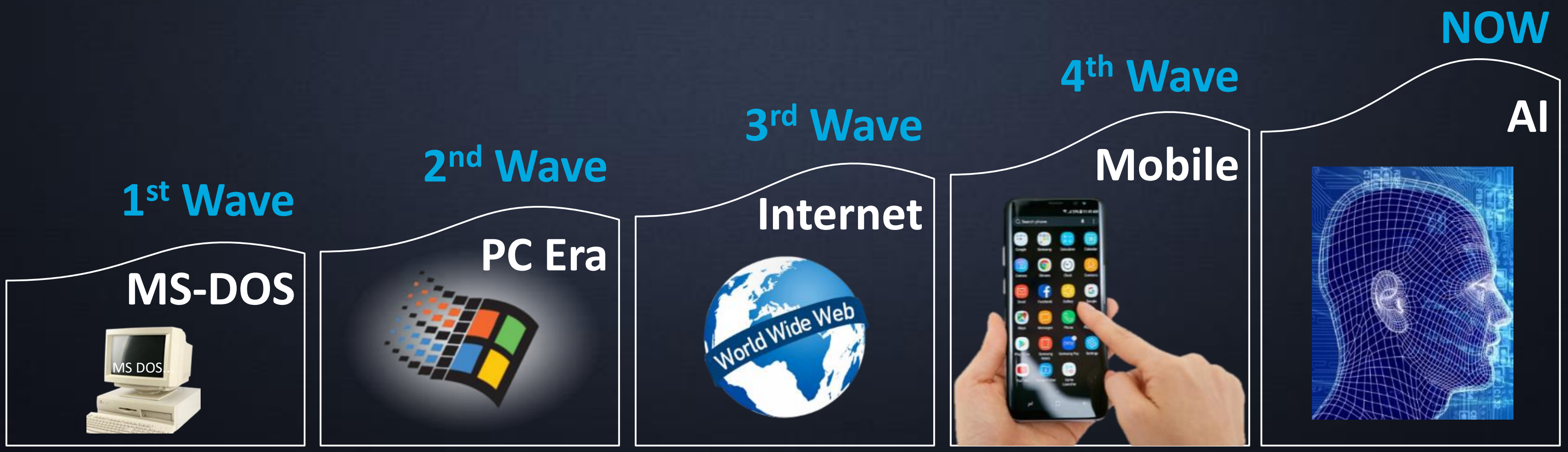
x86 Processors

CPU-centric

Apps Processors

FPGA's
GPU/TPU
Non-x86 processors & platforms

Data-centric



1st Wave
MS-DOS

2nd Wave
PC Era

3rd Wave
Internet

4th Wave
Mobile

NOW
AI

Artificial Intelligence → MAINSTREAM

Speech, Natural Language

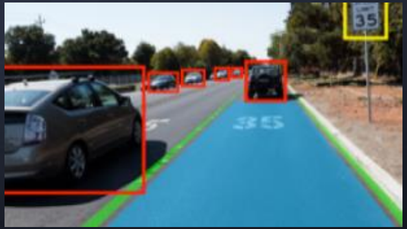
Amazon Echo & Alexa
Google Smart Home Devices
Siri & Cortana Smart Assistants

Deep Learning

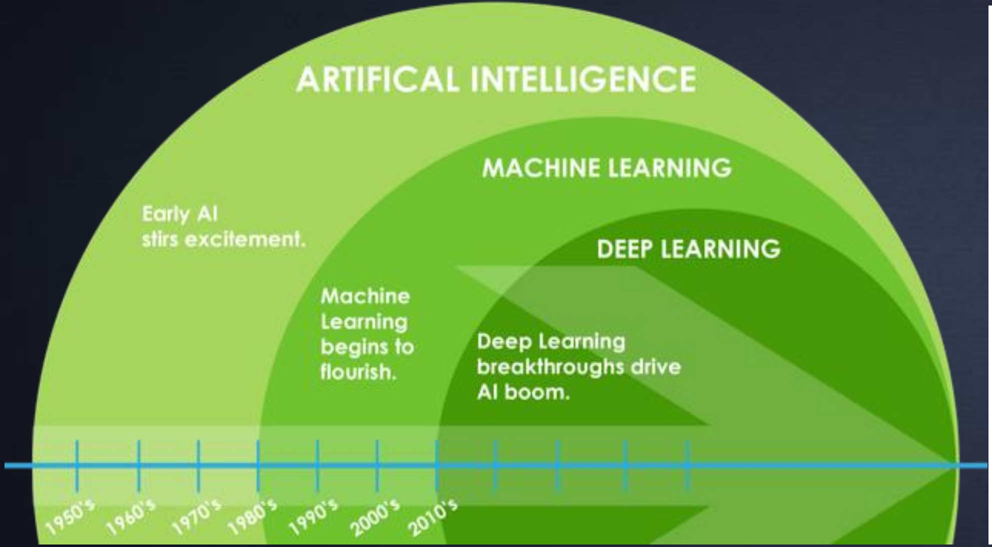
Screening *Genomics* *Prediction* *Game Theory*

Image / Facial Recognition

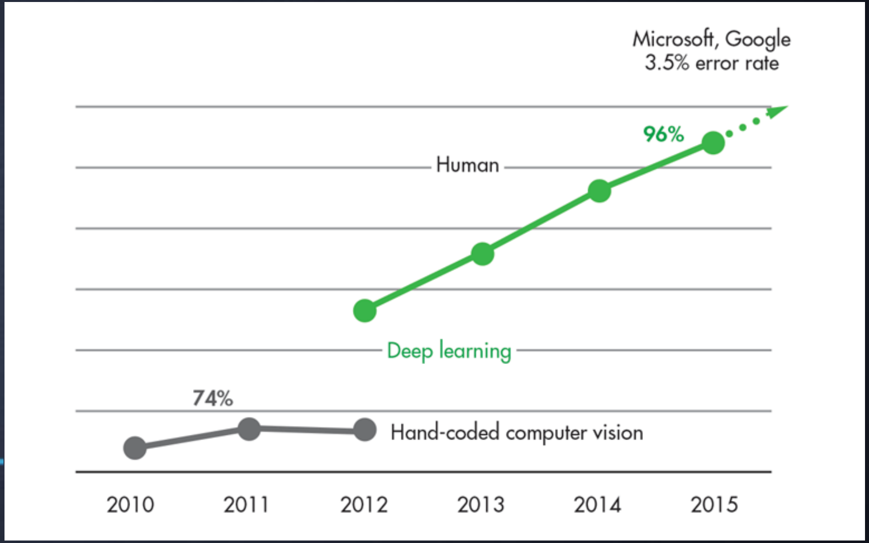
Autonomous Driving



AI – What has Changed?


ARTIFICAL INTELLIGENCE
MACHINE LEARNING
DEEP LEARNING
Early AI stirs excitement.
Machine Learning begins to flourish.
Deep Learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Source: Tuples Edu, buzzrobot.com

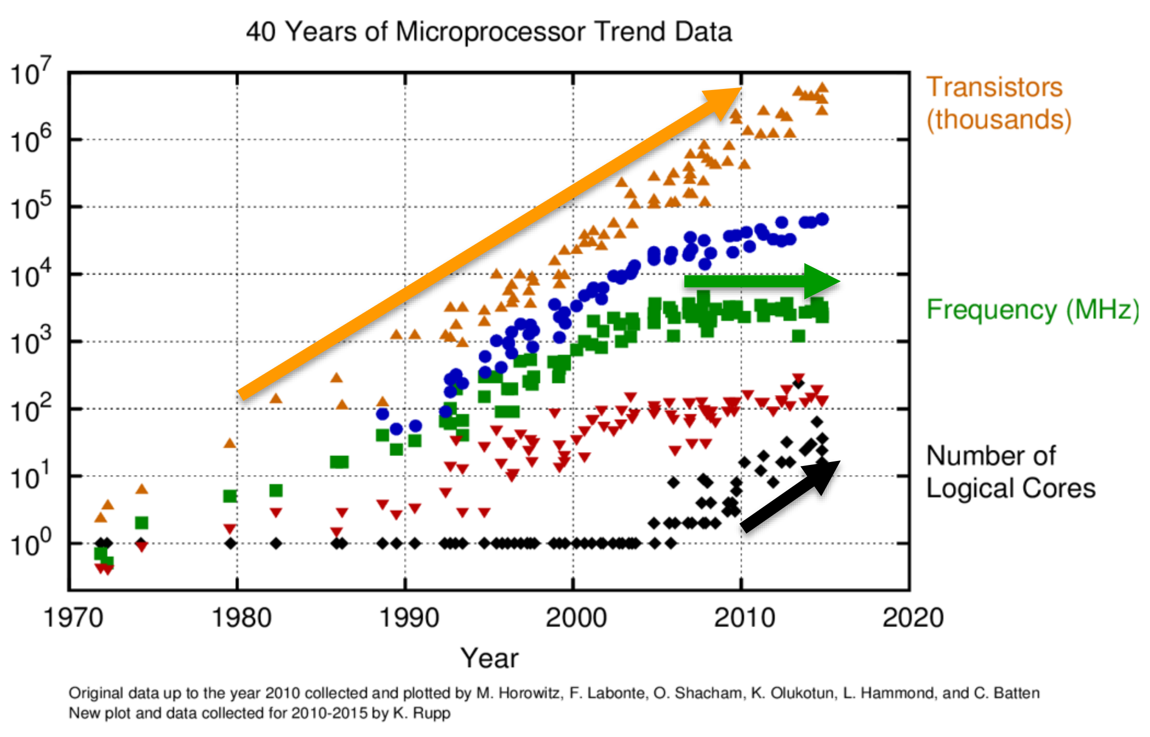


Source: Nvidia, FMS 2017

Deep Learning algorithms require <u>high memory bandwidth</u>

SAMSUNG

Faster Computation → Multi-core

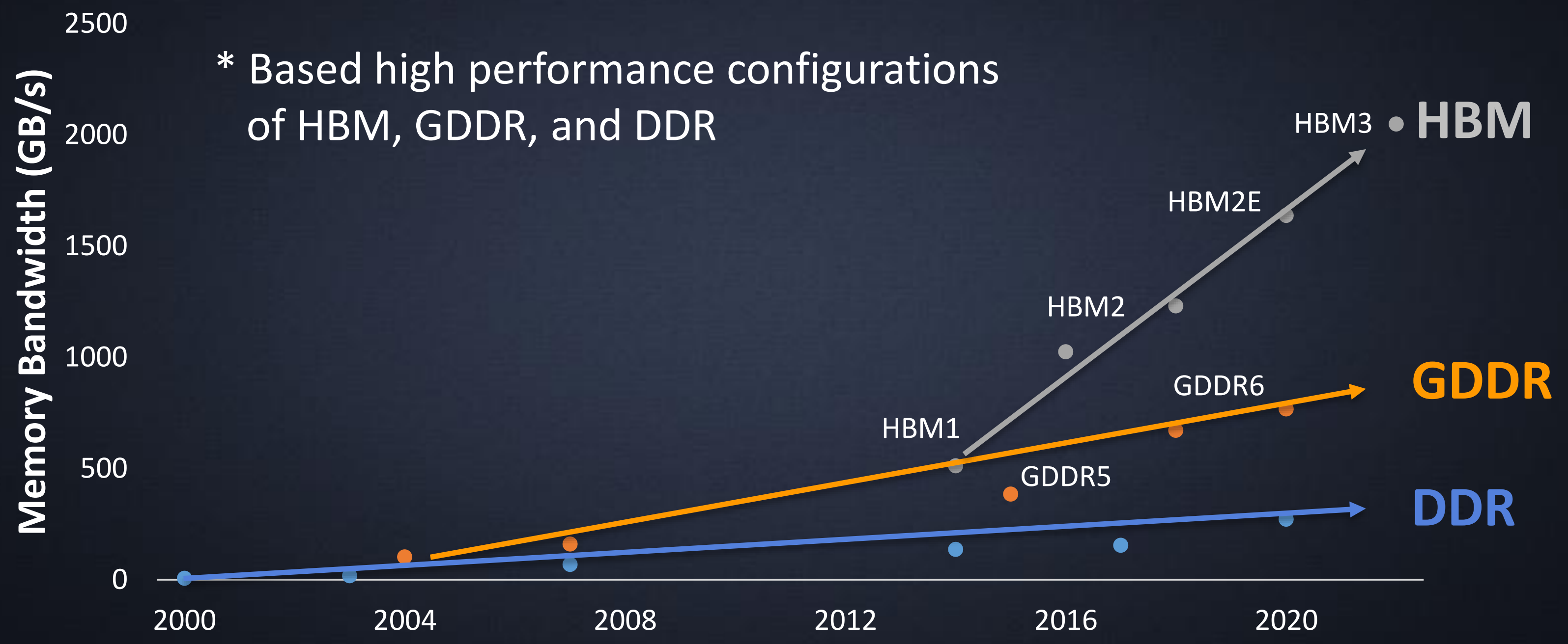


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

High performance compute requires <u>high memory bandwidth</u>

SAMSUNG

Memory Bandwidth Comparison



* Based high performance configurations of HBM, GDDR, and DDR

SAMSUNG

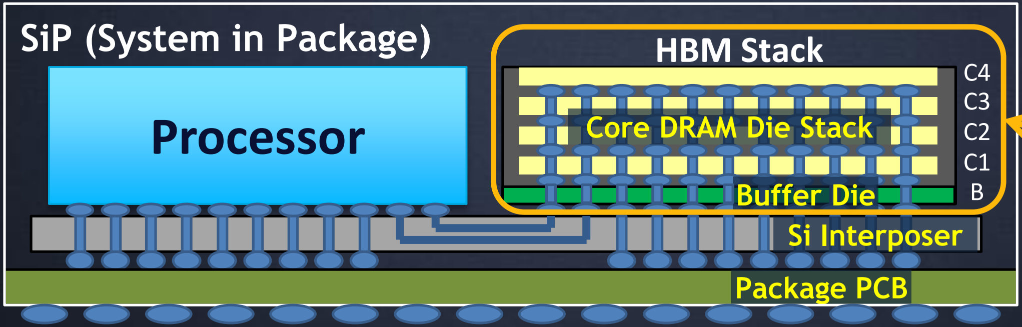
HBM: High Bandwidth Memory

- Stacked MPGA (micro-pillar grid array) memory solution for high performance applications
- Samsung launched HBM2 in Q1 2016
- Uses DDR4 die with TSV (Through Silicon Vias)
- Available in 4H or 8H stacks
- Key Features:
  - 1024 I/O's (8 Channel, 128bits per channel)
  - Per stack: 307GB/s (current generation)
    - 77X the speed of a PCIe 3.0 x4 slot, or
    - 77 HD movies transferred per second

**** Announced HBM2E: +33% throughput (410GB/s), 2X density (16GB stack) ****

HBM Basics: 2.5D System In Package

- A typical HBM SiP consists of a processor (or ASIC) and 1 or more HBM stacks mounted on a Silicon Interposer
- The HBM consists of 4 or 8 DRAM die mounted on a buffer die



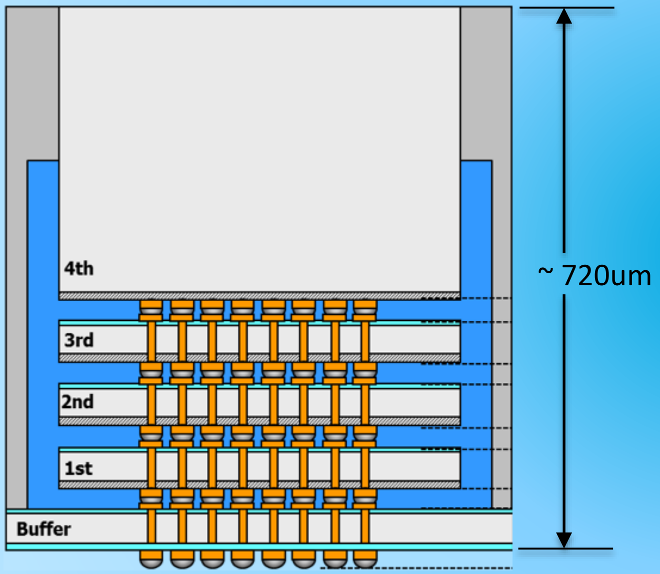
Samsung manufactures and sells the HBM stack

- The entire system (Processor + HBM stack + Si Interposer) is encapsulated into one larger package by the customer

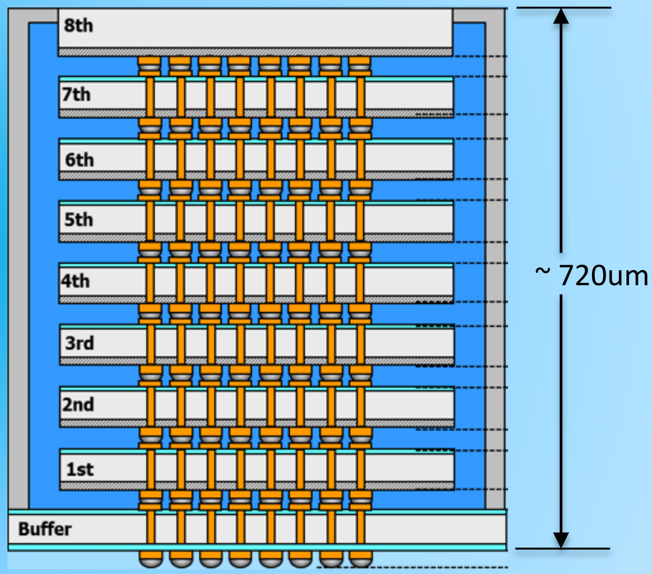
SAMSUNG

MPGA: Micro-Pillar Grid Array

Four High Stack (4H)

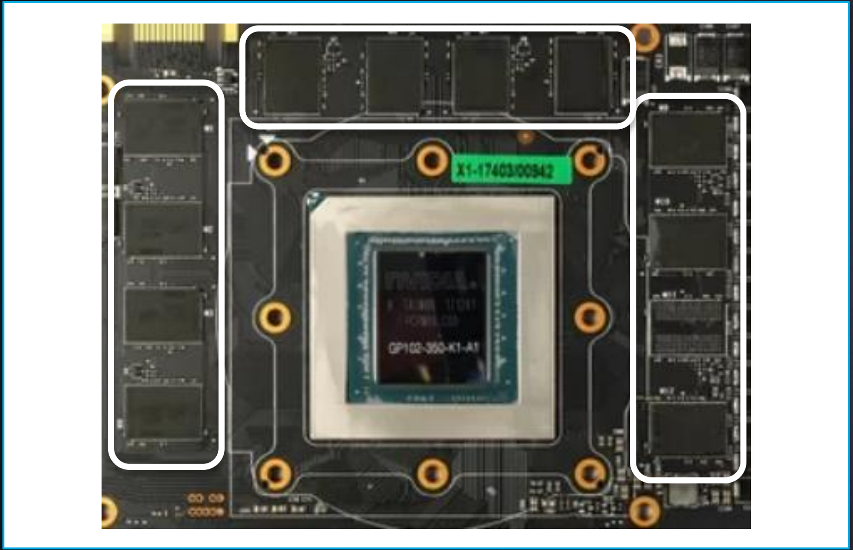


Eight High Stack (8H)



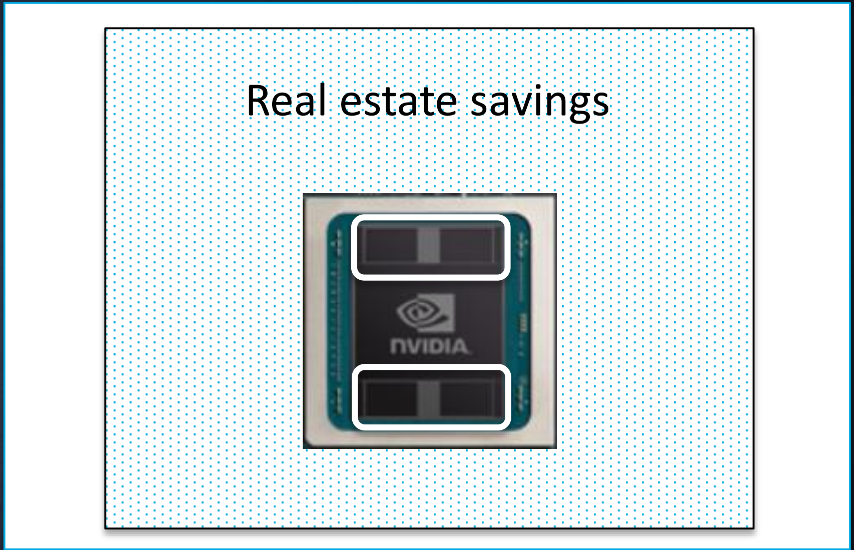
Not just about speed: Space Efficiency

GDDR5



Density	1 GB x 12 = 12GB
Speed/pin	1 GB/s
Pin count	384
B/W	384 GB/s

HBM2E

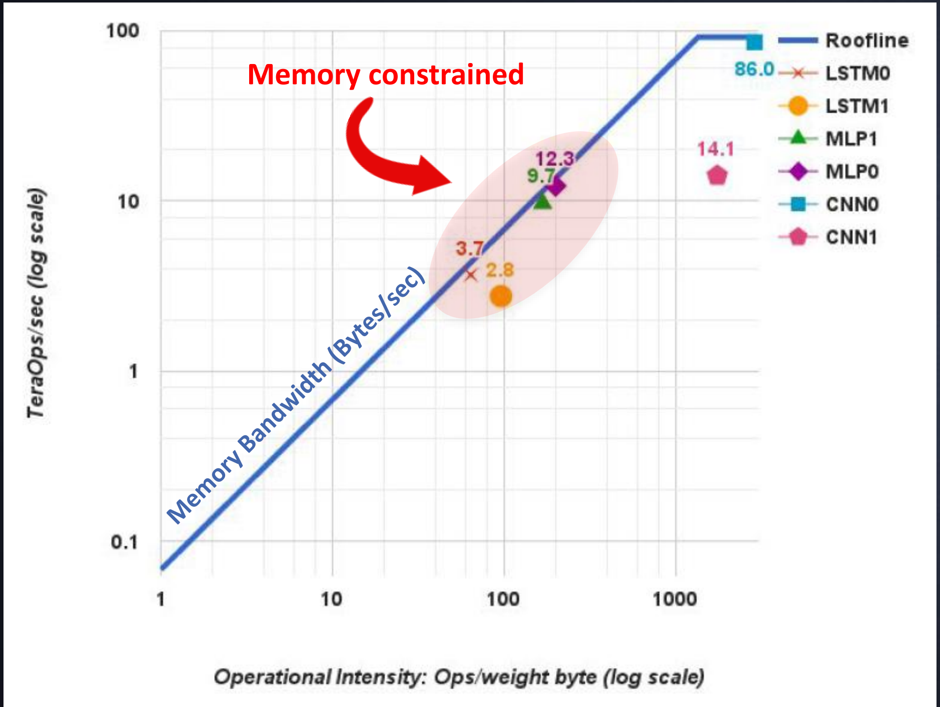


Real estate savings

Density	16 GB x 4 = 64GB
Speed/pin	0.4 GB/s
Pin count	4096
B/W	1,640 GB/s

AI: Compute vs. Memory Constrained

Roofline Model for TPU ASIC



Source: Google ISCA 2017

Roofline Model

- Point below slope = memory bandwidth constrained
- Point below horizontal = compute constrained

Neural Network	Characteristic	Use Case
MLP	Structured input features	Ranking
CNN	Spatial processing	Image recognition
RNN	Sequence processing	Language translation

* LSTM (Long Short-Term Memory) is subset of RNN

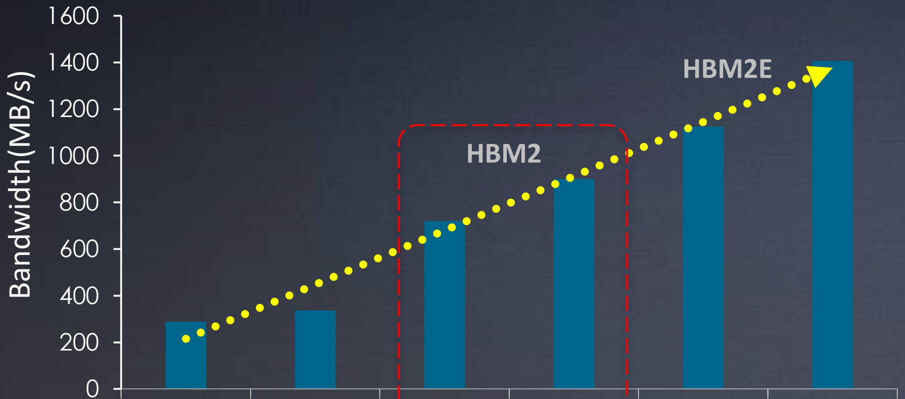
Many Deep Learning applications are MEMORY bandwidth constrained
*→ Need **High Bandwidth Memory***

Memory Drives AI Performance

✓ Faster Training, More Bandwidth

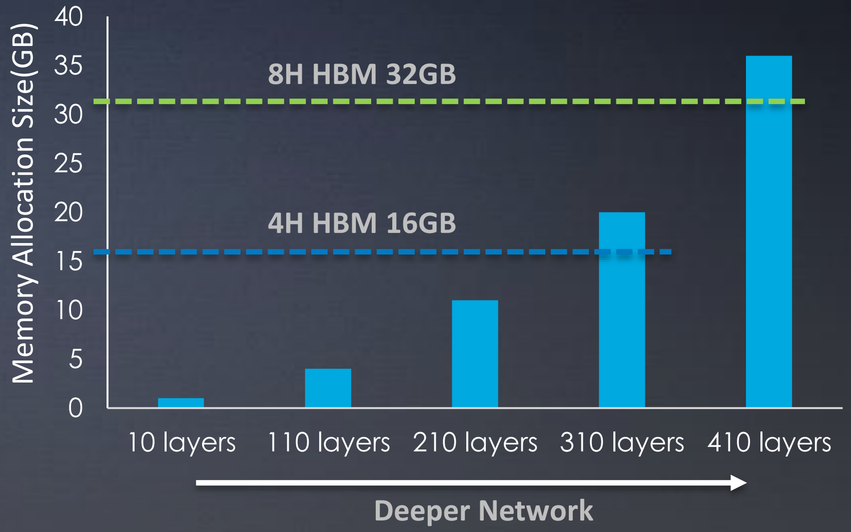
✓ Better Accuracy, More Capacity

Required Memory BW (GB/s)



TFLOPS,	5.2	7	10	15	?	?
# Core,	2880	3072	3584	5120		
Product	K110	M200	P100	V100		

Memory allocation size (GB)



HBM Presence – Some Examples

NVIDIA

Datacenter (Acceleration, AI/ML)
- Tesla P100, V100
- DGX Station, DGX1, DGX2
- GPU Cloud
- Titan V

AI Cities
Healthcare
Retail
Robotics
Autonomous cars

Professional Visualization
- Quaddro GP100, GV100

Architecture
Engineering/Construction
Education
Manufacturing
Media & Entertainment

AMD

Datacenter (Acceleration, AI/ML)
- Radeon Instinct MI25
- Project 47

Traffic sign recognition
Image synthesizer
Object classifier
Model conversion

Professional Visualization
- Radeon Pro WX, SSG, Vega

VR content creation
Graphics rendering

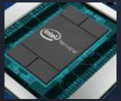
Consumer Graphics
- Radeon Rx Vega64, Vega56

Gaming, AR/VR

intel

Datacenter (Acceleration, AI/ML)
- Nervana Neural Net Processor
- Stratix10 MX (FPGA)

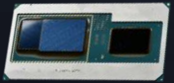
ASIC



FPGA

Consumer Graphics
- KabyLake-G

CPU/GPU Hybrid



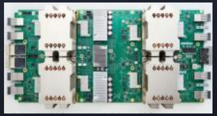
H/E GFX in notebooks
Thin/light
Extended battery life

Google

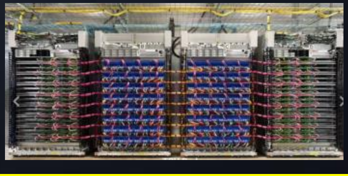
Datacenter (Acceleration, AI/ML)
- TPU2

Cloud TPU for Training & Inference

TPU2: 4 ASICs, 64GB HBM2



TPU POD: 4TB HBM2

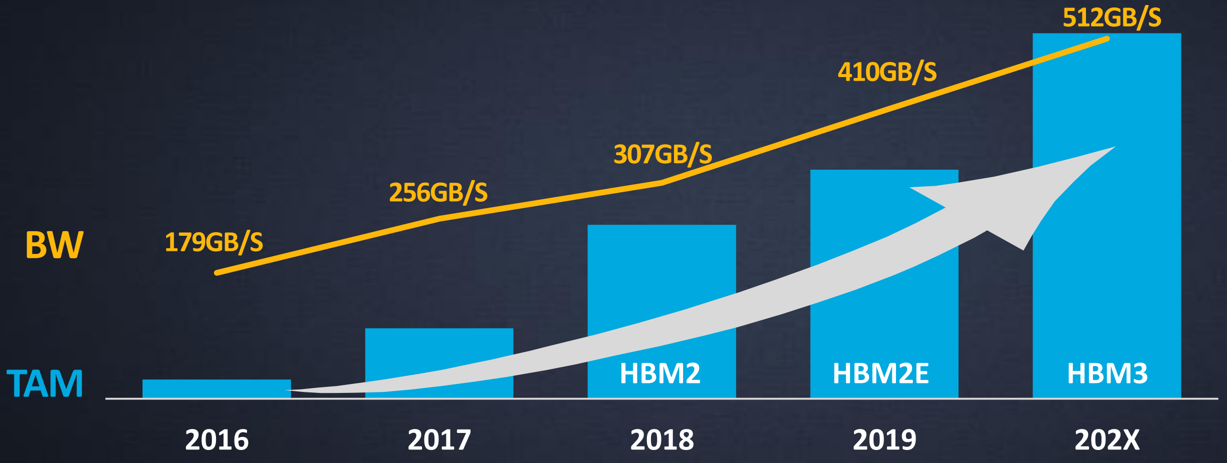


SAMSUNG

Sources: Tom's Hardware, Anandtech, PC World, Trusted Reviews

HBM2: Market Outlook

- Bandwidth needs of High-Performance Computing/AI, High-end Graphics, and new applications continue to expand



Bandwidth and market for HBM growing rapidly

Applications	2016	2017	2018	2019	202X
	HPC/AI	HPC/AI	HPC/AI	HPC/AI	HPC/AI
		Networking	Networking	Networking	Networking
		VGA	VGA	VGA	VGA
			Others	Others	Others

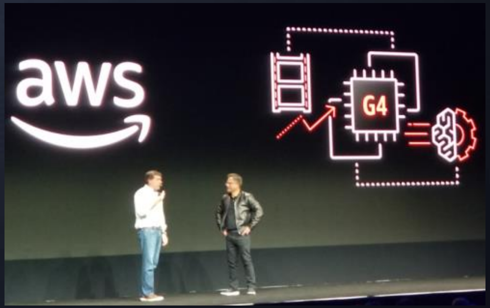
HBM adoption started with HPC, expanding into other markets

Source: Samsung

SAMSUNG

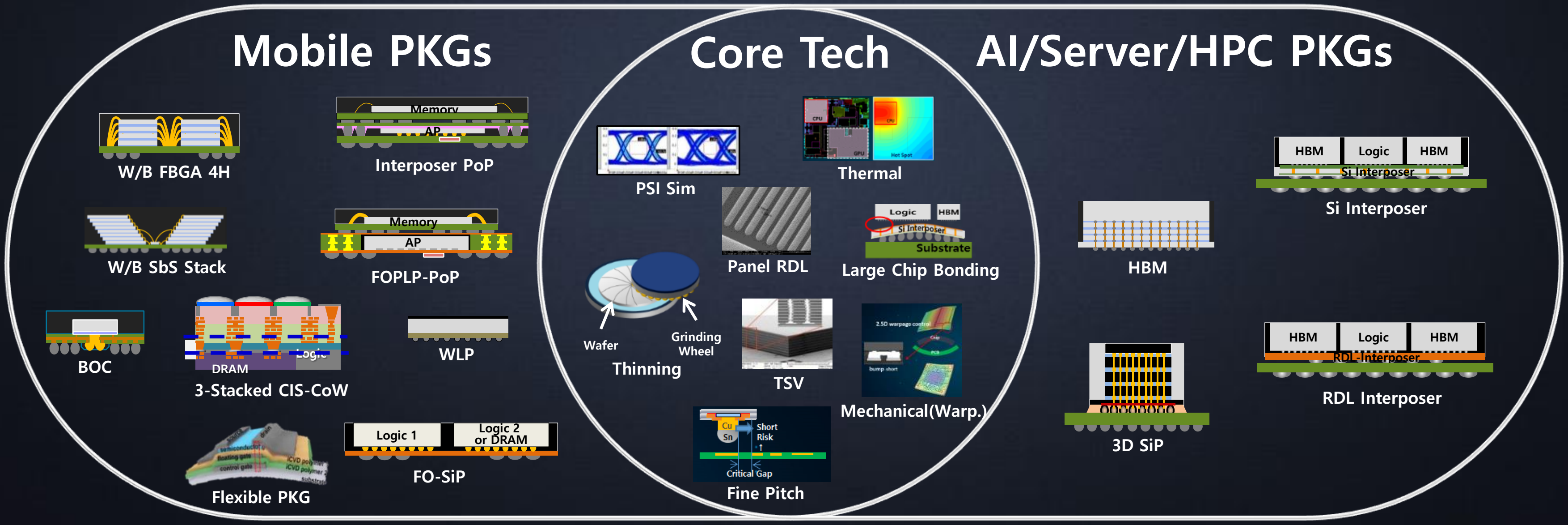
AI Inference: GDDR6

- Inference less computationally & memory intensive than AI Training
- GDDR6 is a good option – double the bandwidth of GDDR5
  - Up to 16Gbps per pin → 64GB/s per device
- Samsung is first to market with 16Gb GDDR6
- Nvidia T4 cards
  - 16GB GDDR6
  - AWS G4 Inference



Foundry Services

- Latest process nodes, testing, packaging, design services
- WW partners to complement solutions with IP and EDA tools



Mobile PKGs

W/B FBGA 4H

Interposer PoP

W/B SbS Stack

FOPLP-PoP

BOC

3-Stacked CIS-CoW

WLP

Flexible PKG

FO-SiP

Core Tech

PSI Sim

Thermal

Panel RDL

Large Chip Bonding

Thinning

TSV

Mechanical(Warp.)

Fine Pitch

AI/Server/HPC PKGs

Si Interposer

HBM

3D SiP

RDL Interposer

Summary

- AI workloads rely on Deep Learning algorithms that are memory bandwidth constrained

- HBM has become the memory of choice for AI training applications in the data center

- GDDR6 provides an “off-the-shelf” alternative for AI inference workloads

Make the smart choice: AI hardware powered by these technologies

Thank You…

Samsung@
The Heart of
Your Data

March 19-21, 2019

SAN JOSE CONVENTION CENTER · BOOTH 726

Contact: t.shiah@Samsung.com