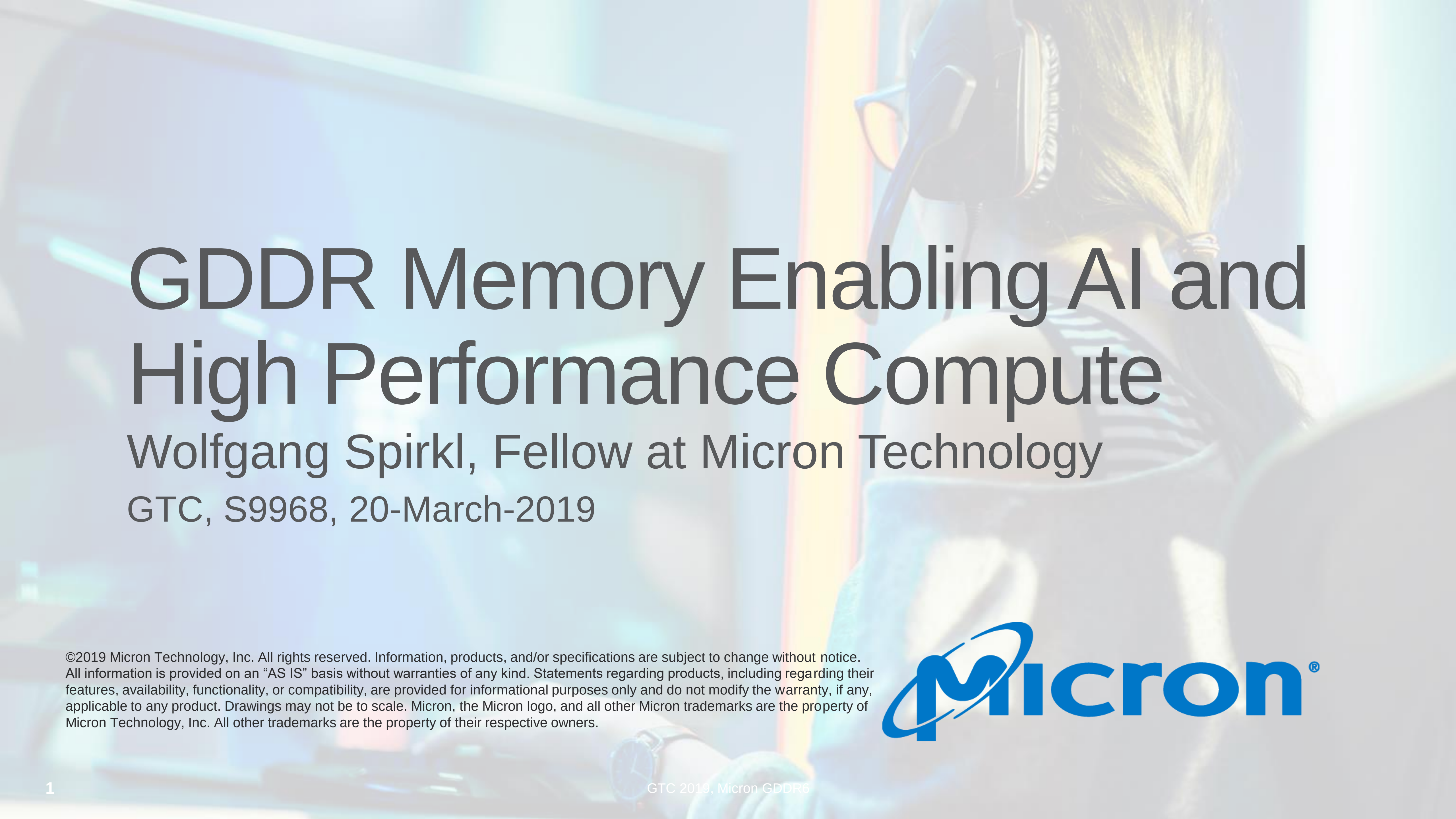


GDDR Memory Enabling AI and High Performance Compute

Wolfgang Spirkl, Fellow at Micron Technology

GTC, S9968, 20-March-2019

Agenda

- The Demand for faster Memory and storage
- Competing Compute/Memory Solutions
- GDDR6 for AI applications and more
- Micron GDDR6 AI demonstration

Accelerated Data Cycle

Driven by Increasing Data Value

- Creates continuous need to capture, process, move & store data
- Generates ever-increasing demand for memory & fast storage
- AI is amplifying the Accelerated Cycle

AI Landscape for Memory & Storage

Data Center

System/Infrastructure:

Cloud

On Prem.

Workloads:

Model Development

Batch Training

Batch Inference

Memory/Storage:

PERFORMANCE

HBM | P-Mem

GDDR6 | 3D TLC

DDR4 | 3D QLC

Smart Edge/Intelligent Endpoint

System/Infrastructure:

Edge Compute

Smart Access Point

Autonomous Vehicle/Robot

Mobile Device

Smart IoT/ Sensor

Workloads:

Online Training

On-demand Inference

Memory/Storage:

HBM | P-Mem

GDDR6 | 3D TLC

DDR4 | 3D QLC

LP4X

Workloads:

Local Inference

01000110
01100001
01110
01110

Data Collection

Memory/Storage:

3D TLC

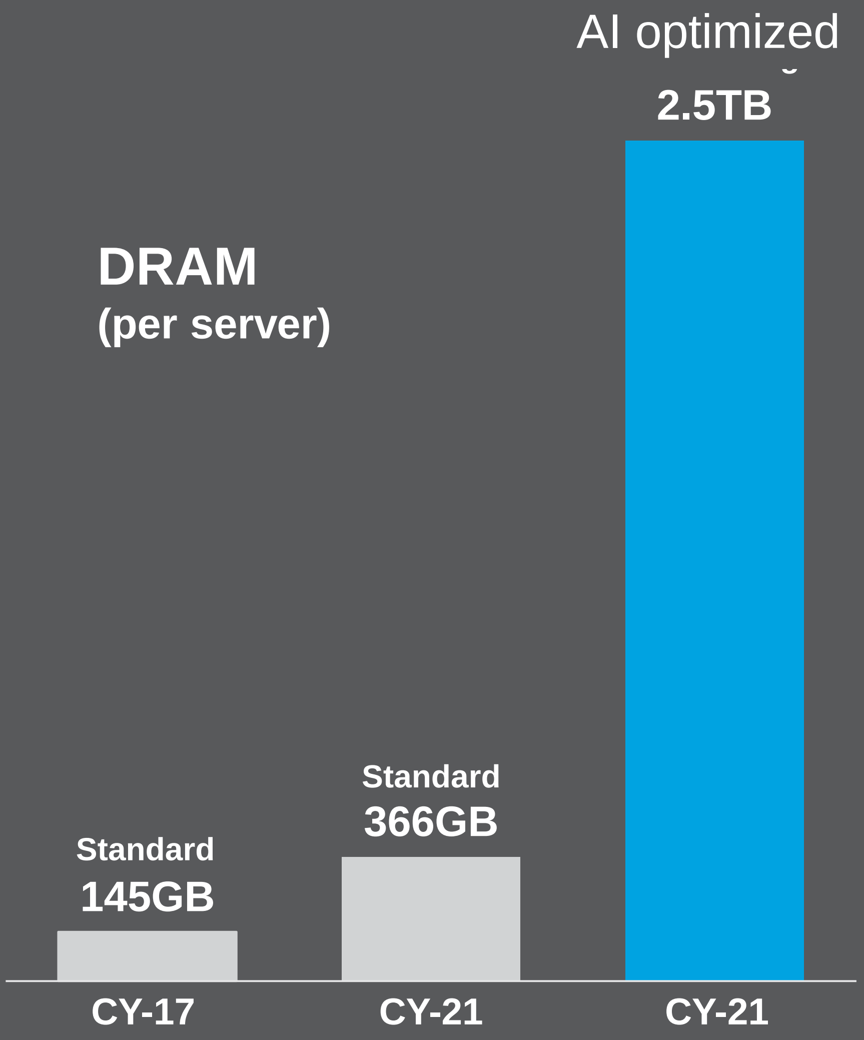
DDR4

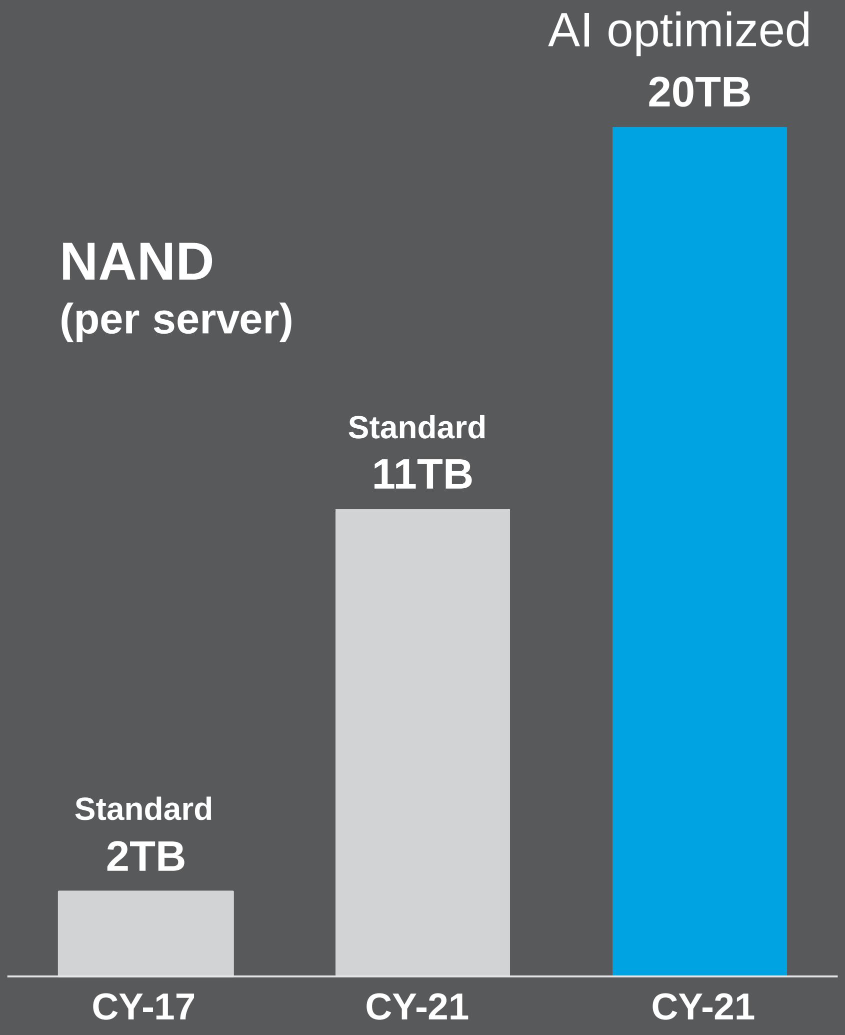
LP4X

COST/POWER

AI Workloads Unleash the Need For More Memory & Storage

Significant Growth Across Private, Public & Hybrid Cloud





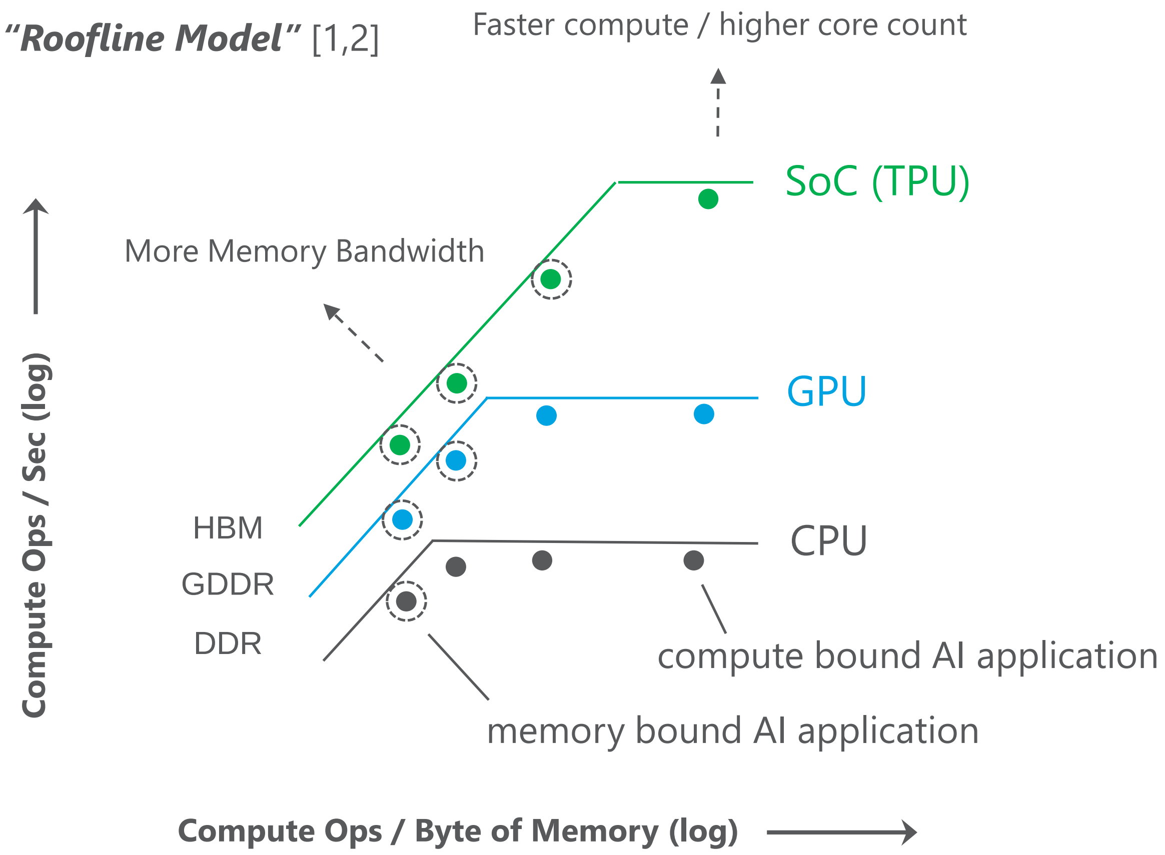
Source: Micron

- The Demand for faster Memory and storage
- Competing Compute/Memory Solutions
- GDDR6 for AI applications and more
- Micron GDDR6 AI demonstration

AI Acceleration is Driving Demand for Memory Bandwidth

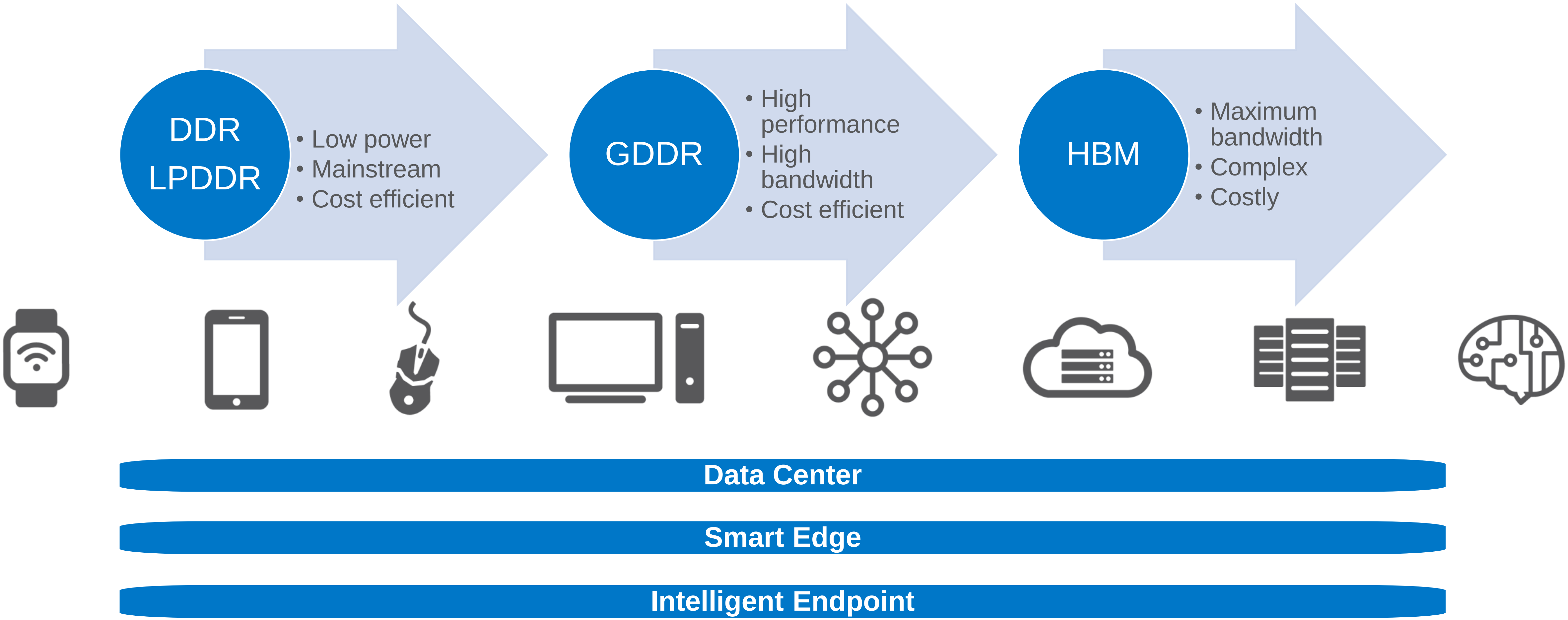
- AI accelerators increase compute performance
  - GPU, TPU, etc..
- Accelerated applications are more likely to be memory bound [3]
- Micron supports next gen technologies
  - GDDR6, HBM2E

"Roofline Model" [1,2]


Faster compute / higher core count
SoC (TPU)
More Memory Bandwidth
GPU
CPU
HBM
GDDR
DDR
compute bound AI application
memory bound AI application
Compute Ops / Sec (log)
Compute Ops / Byte of Memory (log)

[1] Jouppi, Norman, et al, 2017. In-Datacenter Performance Analysis of a TPU, ISCA
[2] Williams, S., Waterman, A. and Patterson, D., 2009. Roofline: an insightful visual performance model for multicore architectures. Communications of the ACM.
(3) Forrester report on memory and storage impact on AI

Memory Options



DDR
LPDDR

- Low power
- Mainstream
- Cost efficient

GDDR

- High performance
- High bandwidth
- Cost efficient

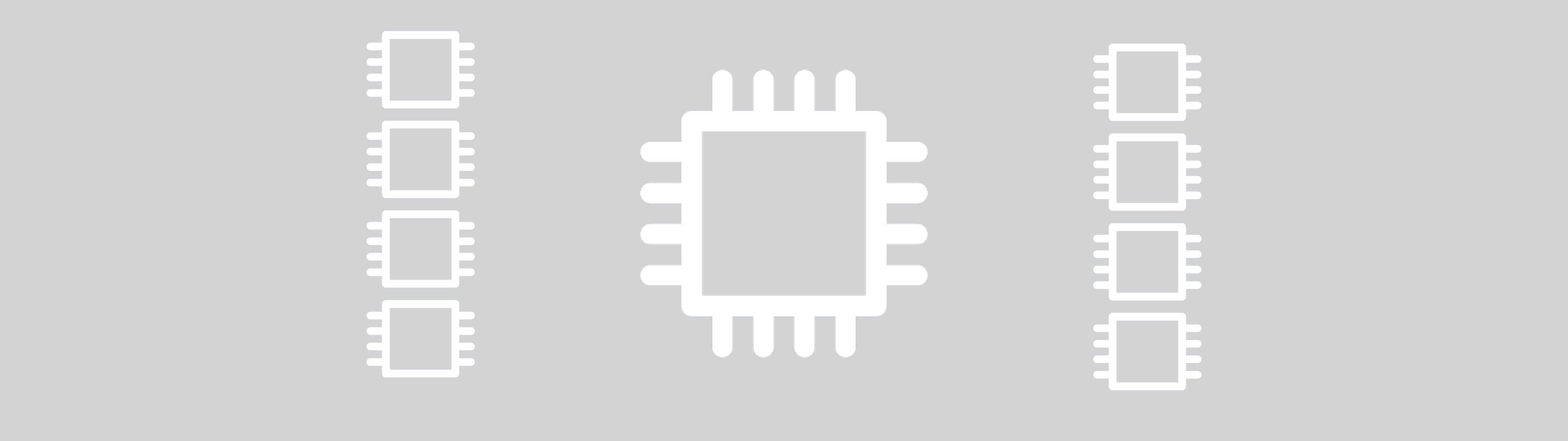
HBM

- Maximum bandwidth
- Complex
- Costly

Data Center

Smart Edge

Intelligent Endpoint



The Demand for faster Memory and storage

Competing Compute/Memory Solutions

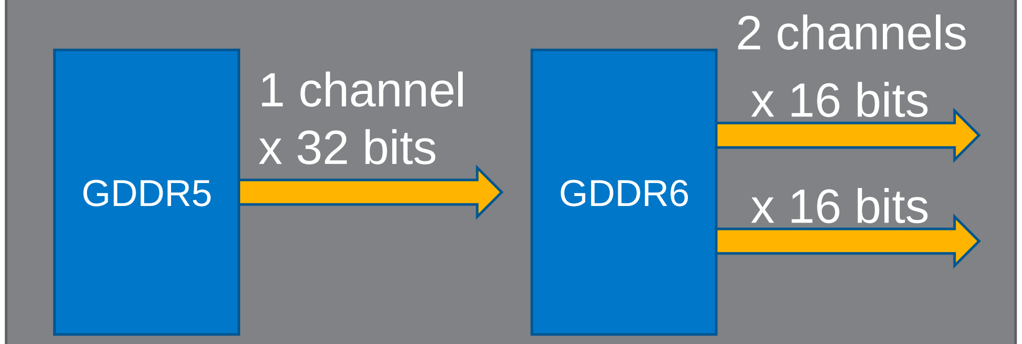
GDDR6 for AI applications and more

Micron GDDR6 AI demonstration

GDDR5/GDDR6 Features

Feature	GDDR5	GDDR6
Density	512Mb – 8Gb	8Gb – 32Gb
VDD, VDDQ	1.5V + 1.35V	1.35V
VPP	N/A	1.8V
Package	BGA-170 14mm x 12mm 0.8mm ball pitch	BGA-180 14mm x 12mm 0.75mm ball pitch
Signaling	POD15 / POD135	POD135
Data rate	≤8 Gbps	≤16 Gbps
I/O Width	x32/x16	2-ch x16/x8
Access Granularity	32B	2-ch 32B each or 1-ch 64B w/ PC mode
I/O Count	61	62 / 74
ABI, DBI	✓	✓
CRC	CRC-8 (BL8)	2x CRC-8 (BL16); compressed 2x CRC-8 (BL8)
RDQS Mode	✓ (BL8)	✓ (BL16)
ODT	✓	✓
V_{REFC}	external	external / internal
V_{REFD}	ext. / int.	internal
Temp Sensor	✓	✓

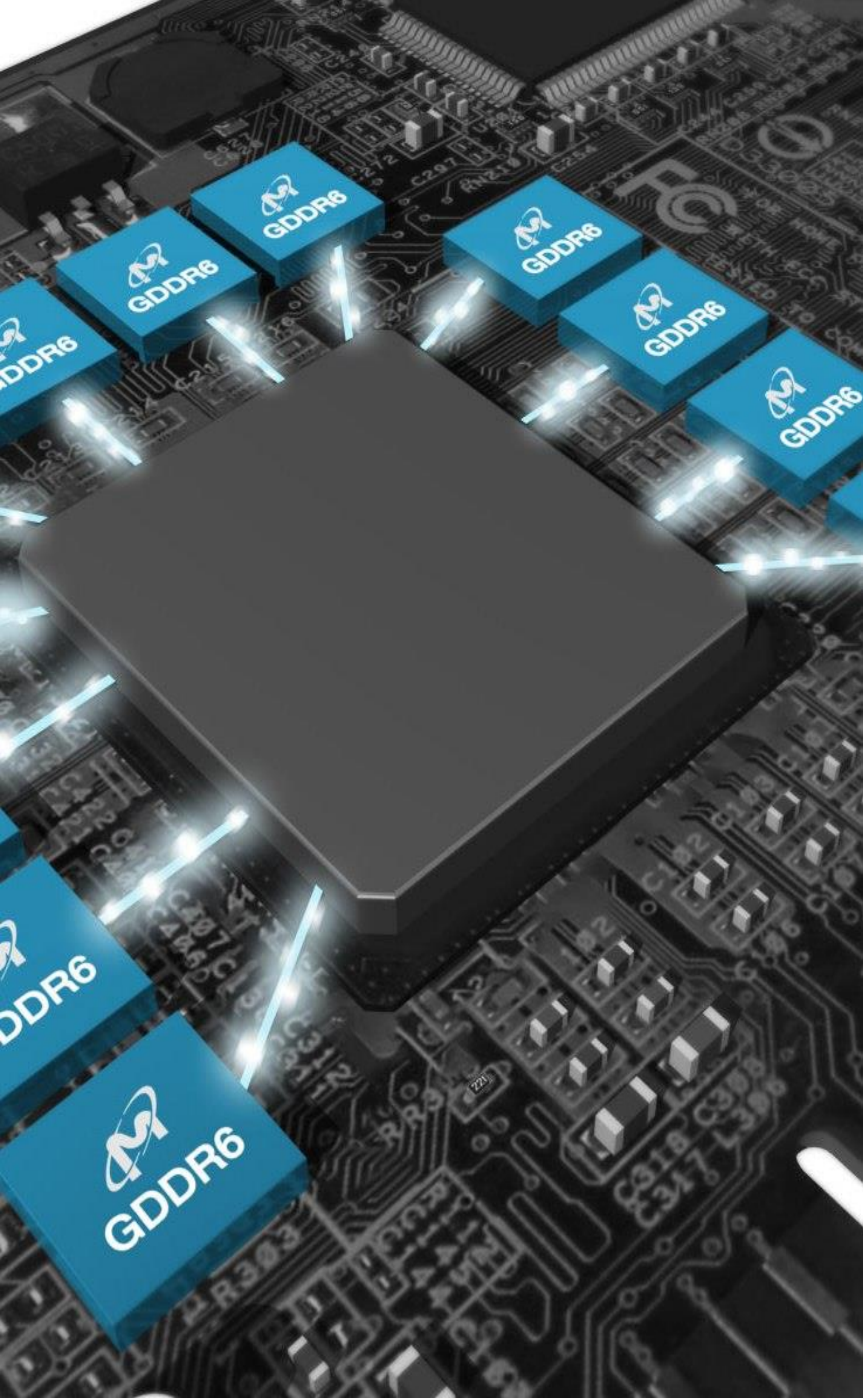
Package Configuration



GDDR6: x16/x8 on dual channel equals GDDR5 x32/x16

Dual channel organization
- Maintains fine granularity (32 bytes per colum access)
- In spite of doubled prefetch size

New features for high data rates:
- Optimized signal ball-out for low-effort PCB design
- Per lane DFE and VREFD
- Transmitter equalization

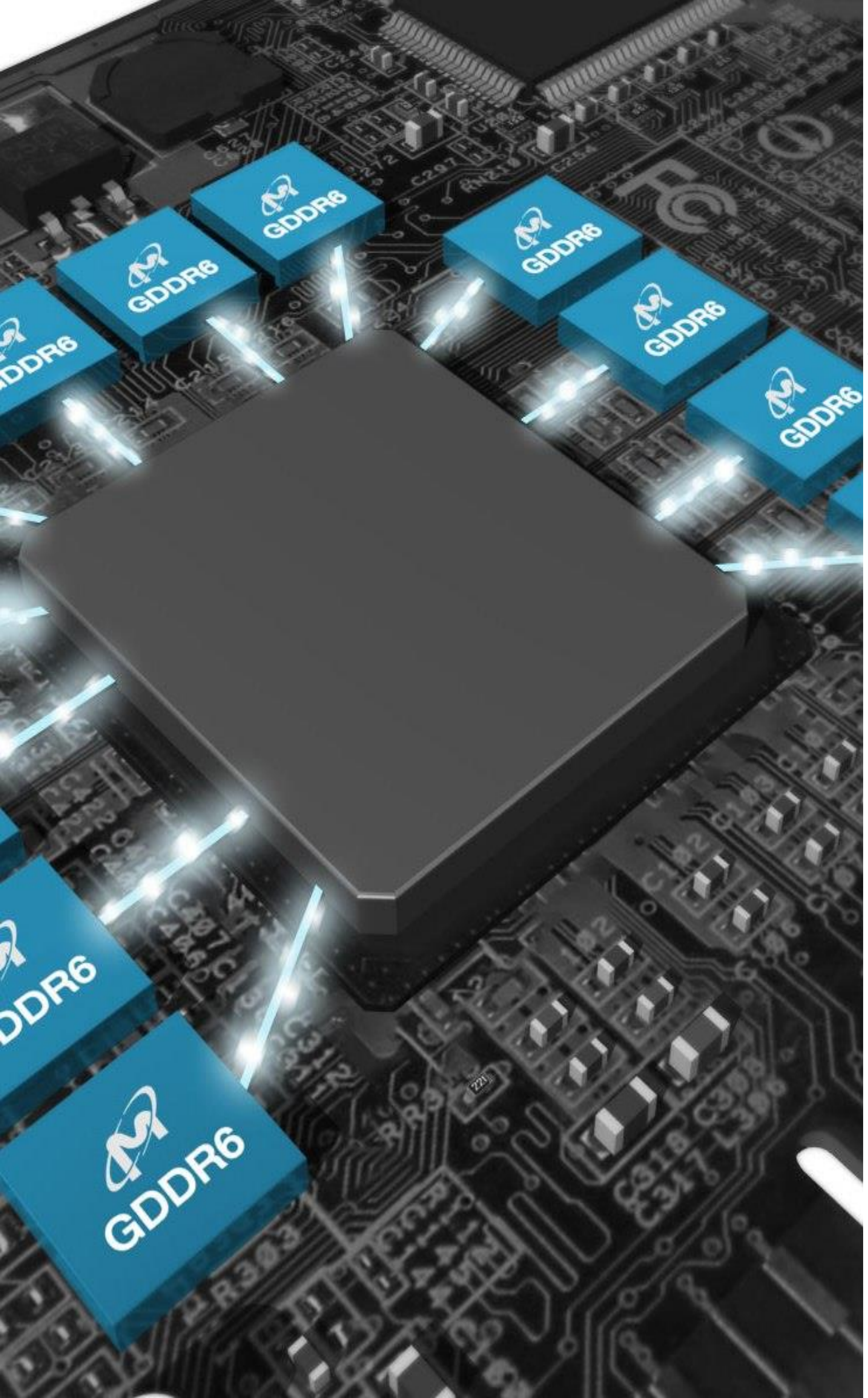


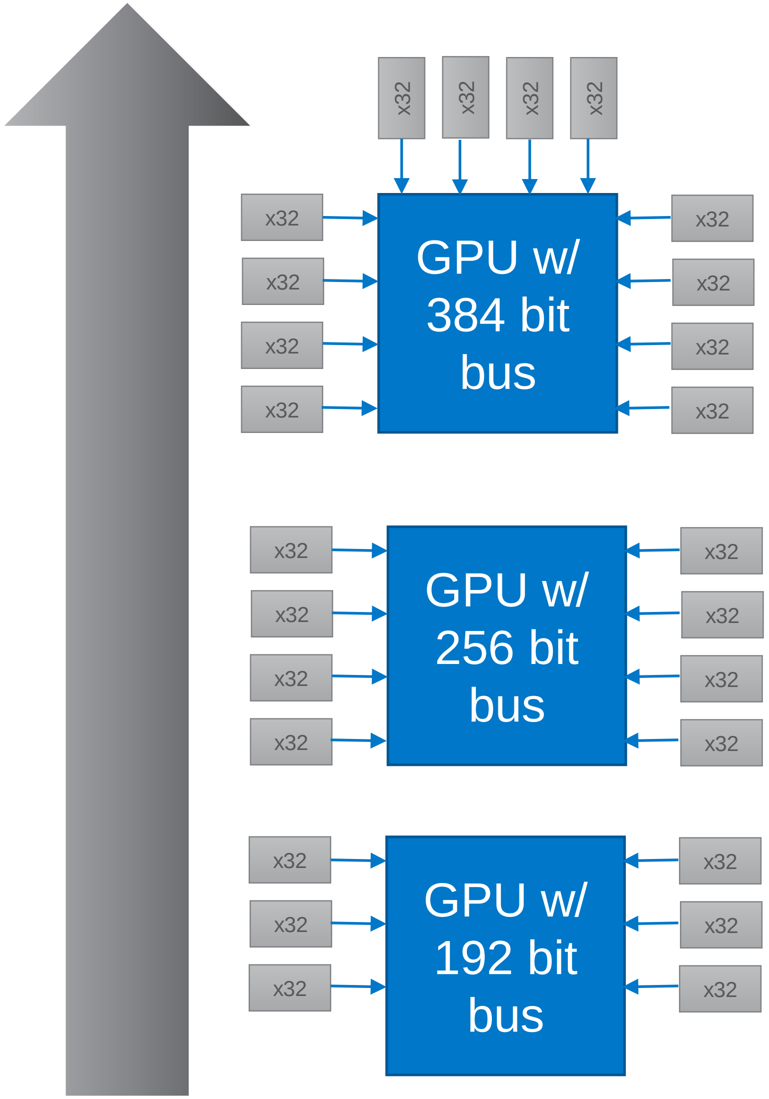
Calculating GDDR Bandwidth

- Bandwidth =
  number of bits/s between GPU and memory
- Memory bus is like traffic lanes
  - More lanes, the greater the flow
  - Higher lane speed, the greater the flow

Memory Bandwidth is	GDDR6 Example
number of memory components X number of lanes per component X Data rate per lane (Gbps)	8 32 16
Memory Bandwidth (GB/s)	512

GDDR Bandwidth / Memory Bus



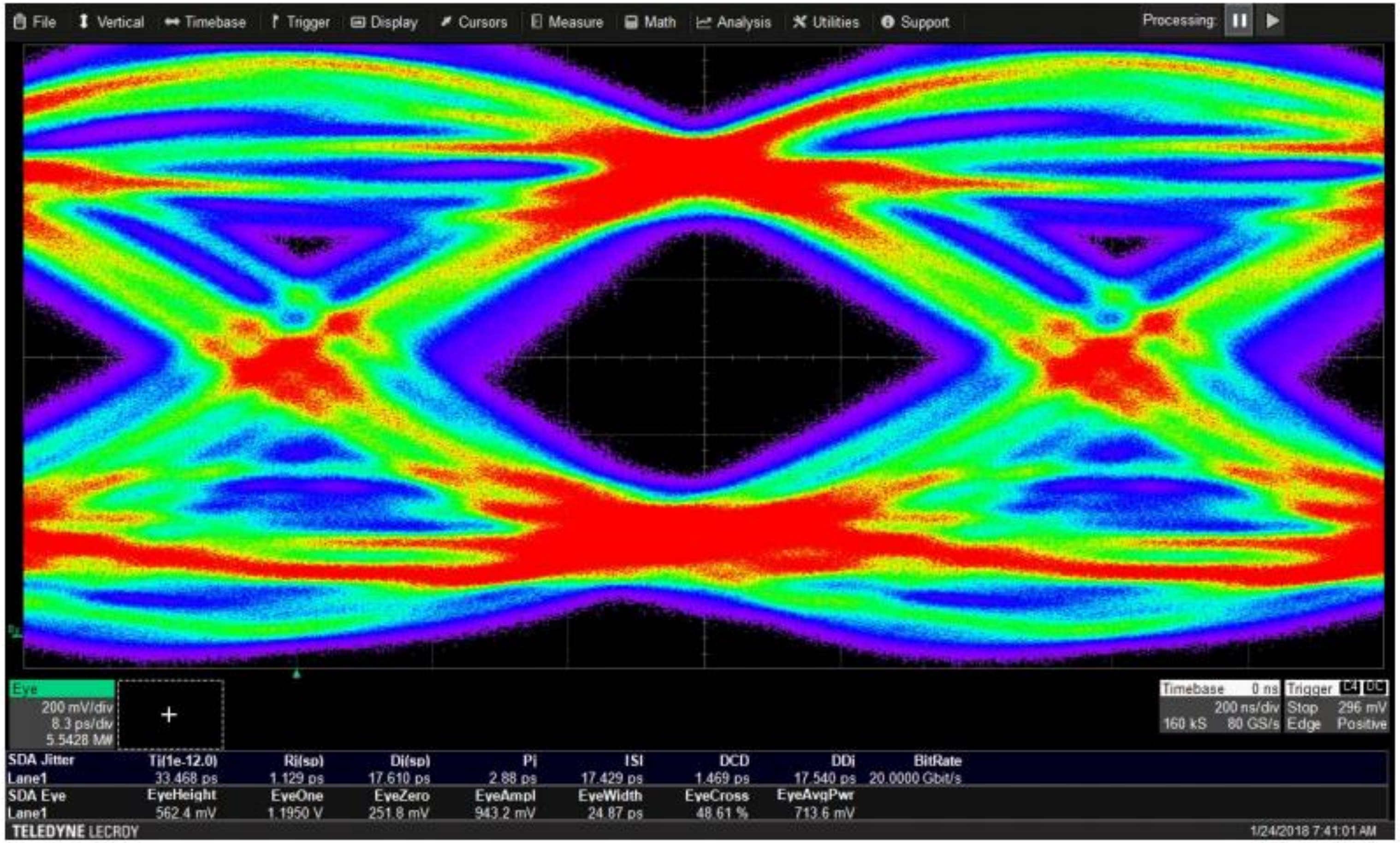


Technology	Speed (Gbps)	# of comp.	# of lanes	Memory bus (bit)	Bandwidth (GB/s)
HBM2	2	4	1024	4096	1024
GDDR6	16	12	32	384	768
GDDR6	14	12	32	384	672
GDDR5X	11	12	32	384	528
GDDR5	7	12	32	384	336
GDDR6	14	8	32	256	448
GDDR5X	11	8	32	256	352
GDDR5	7	8	32	256	224

GTC 2019, Micron GDDR6

GDDR6 20 Gbps Data Eye

Measured performance beyond the specification



Eye
200 mV/div
8.3 ps/div
5.5428 MW

Timebase 0 ns
200 ns/div
160 kS 80 GS/s

Trigger C4 DC
Stop 296 mV
Edge Positive

SDA Jitter	Tj(1e-12.0)	Rj(sp)	Dj(sp)	Pj	ISI	DCD	DDj	BitRate
Lane1	33.468 ps	1.129 ps	17.610 ps	2.88 ps	17.429 ps	1.469 ps	17.540 ps	20.0000 Gbit/s

SDA Eye	EyeHeight	EyeOne	EyeZero	EyeAmpl	EyeWidth	EyeCross	EyeAvgPwr
Lane1	562.4 mV	1.1950 V	251.8 mV	943.2 mV	24.87 ps	48.61 %	713.6 mV

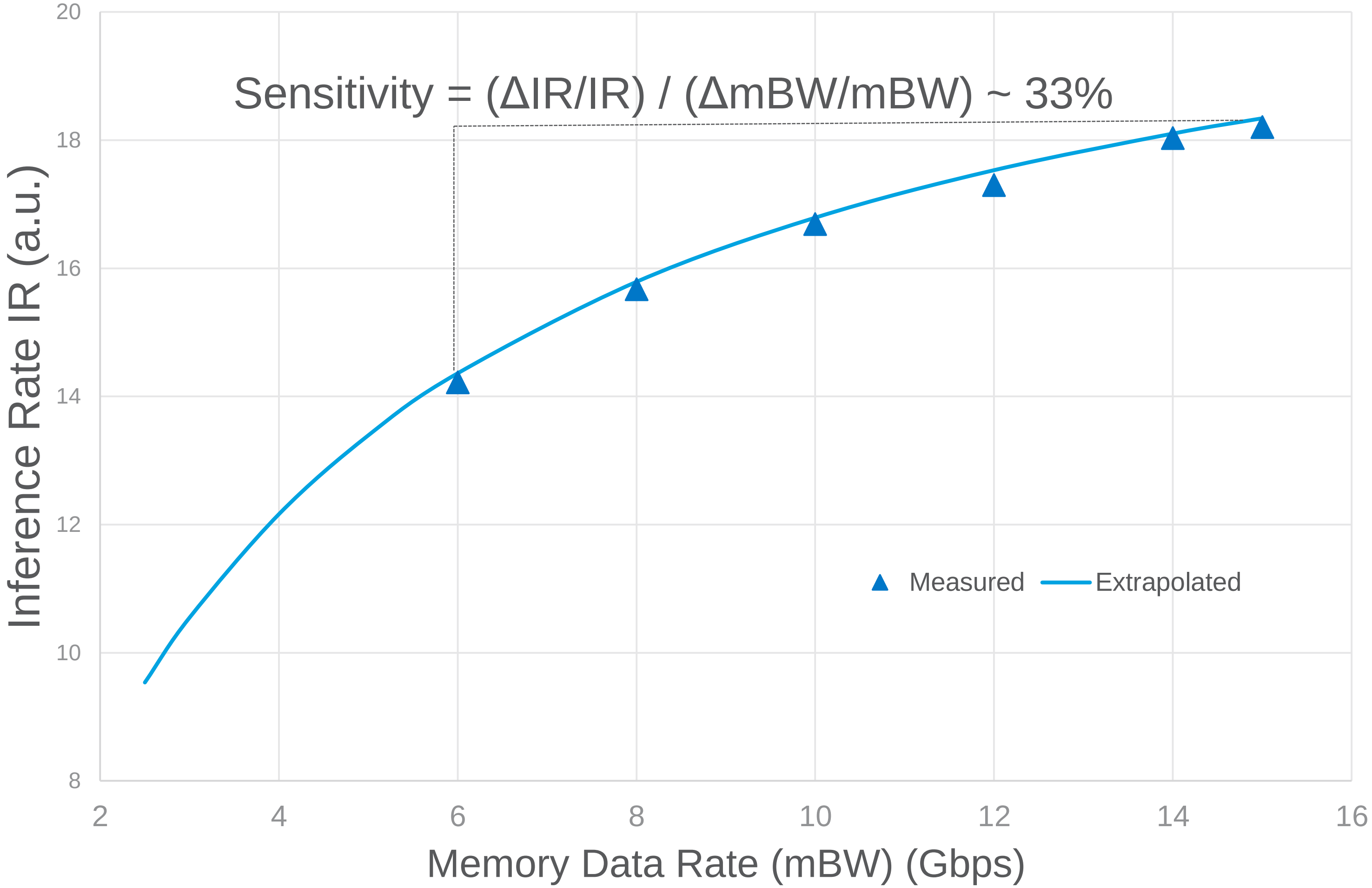
1/24/2018 7:41:01 AM

Figure 15: Measured 20Gb/s data eye based on a PRBS6 pattern

https://www.micron.com/-/media/client/global/documents/products/white-paper/16gb_s_and_beyond_w_single_endedio_in_graphics_memory.pdf?la=en

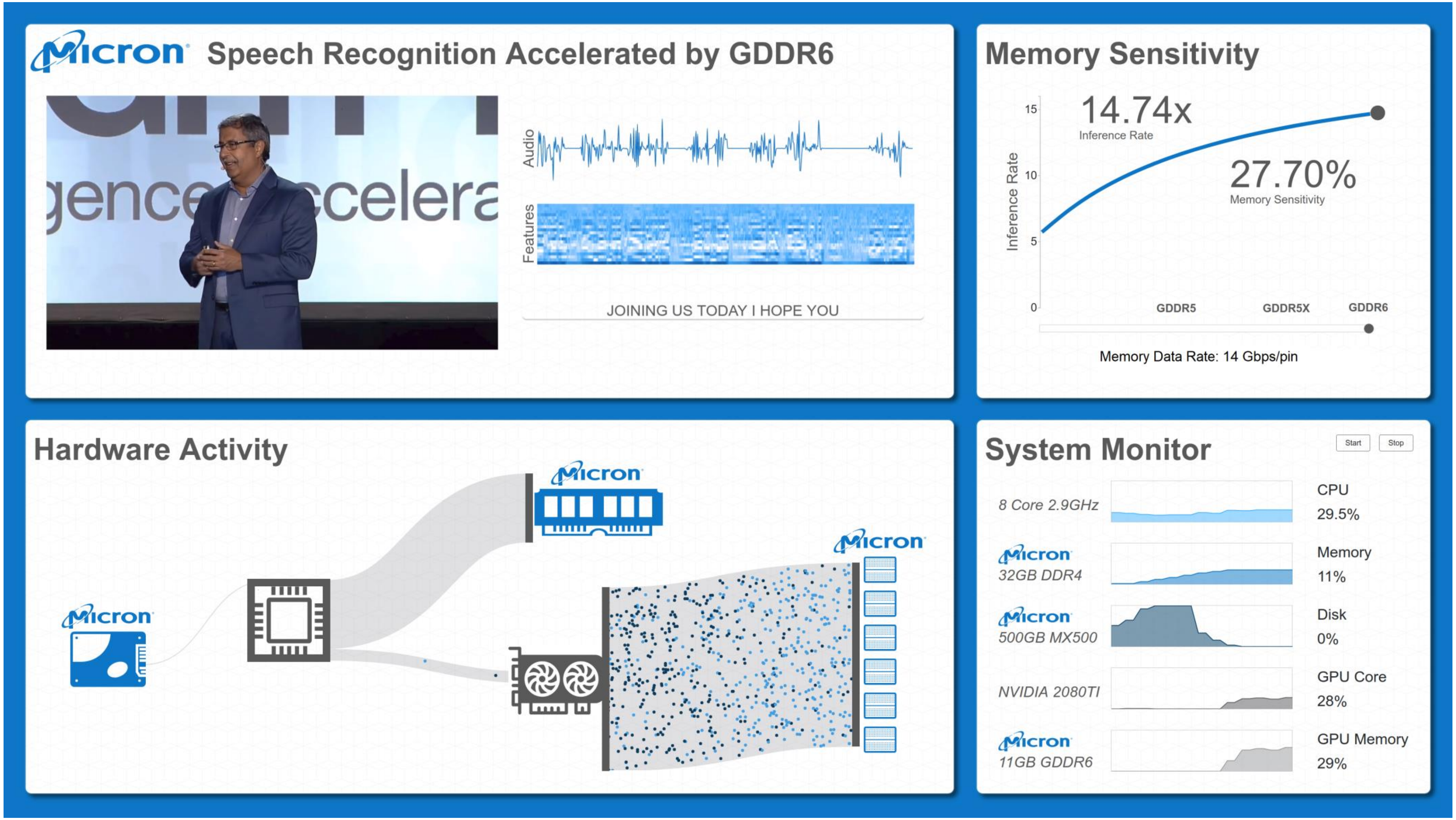
Speech Recognition Craves Memory Bandwidth

- “Deep Speech” recognition application
  - Baidu Research’s AI algorithm (https://arxiv.org/pdf/1412.5567.pdf)
  - Mozilla’s tensorflow implementation
  - Speech-to-text benchmark for AI hardware (https://github.com/mozilla/DeepSpeech)
- Hardware
  - NVIDIA RTX 2080 Ti
  - 11GB GDDR6
    - 384 bit bus @14Gb/s/pin, 672GB/s
- Experiment setup
  - Adjust GDDR6 clock rate
  - Measure speech recognition inference rate:
    - $Inference\ rate = \frac{Audio\ file\ duration}{Inference\ time}$

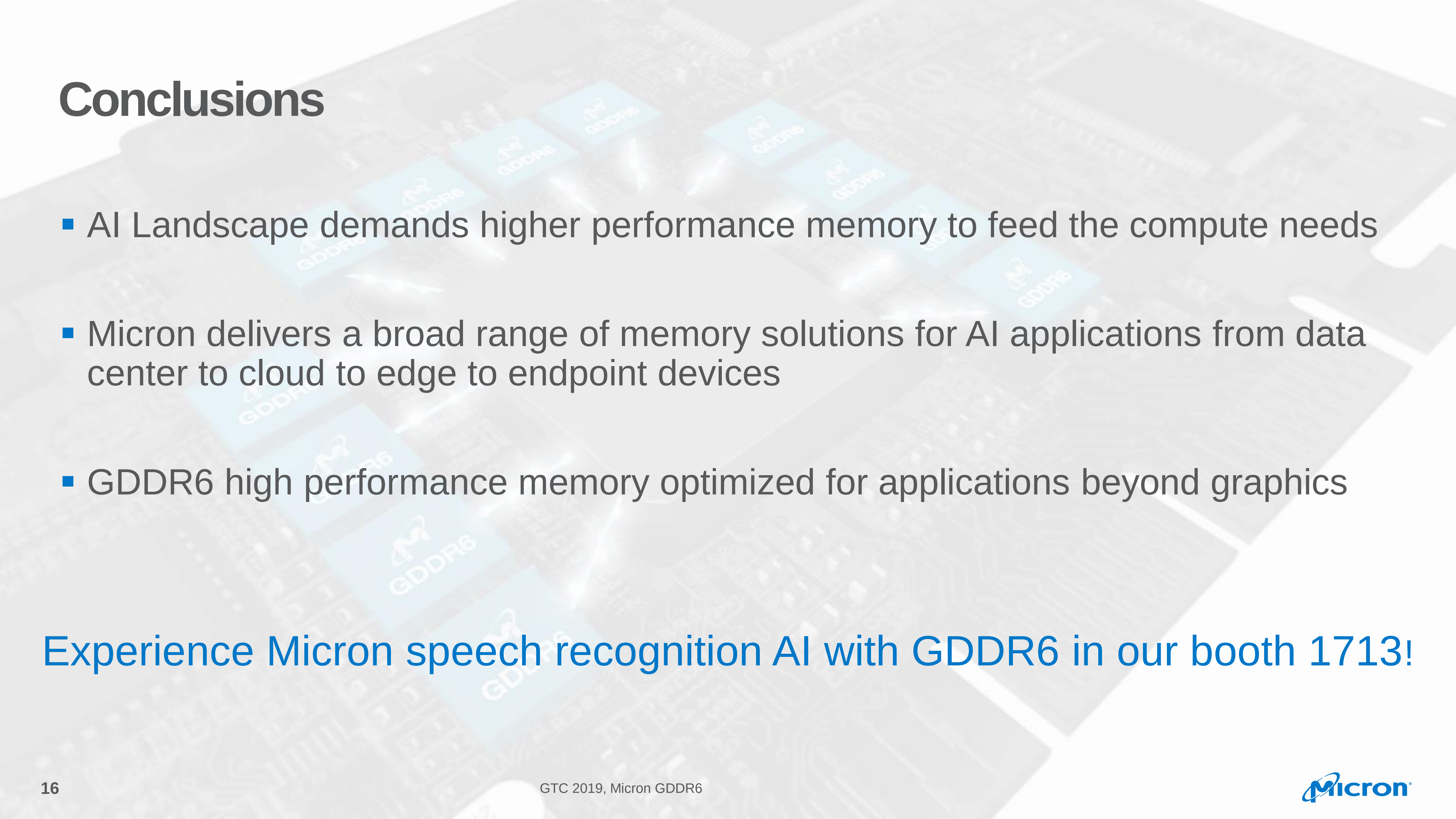


AI Demonstrates the Need for Memory speed

SPEECH RECOGNITION DEMO – Micron Booth # 1713



Conclusions



- AI Landscape demands higher performance memory to feed the compute needs
- Micron delivers a broad range of memory solutions for AI applications from data center to cloud to edge to endpoint devices
- GDDR6 high performance memory optimized for applications beyond graphics

Experience Micron speech recognition AI with GDDR6 in our booth 1713!

Micron®