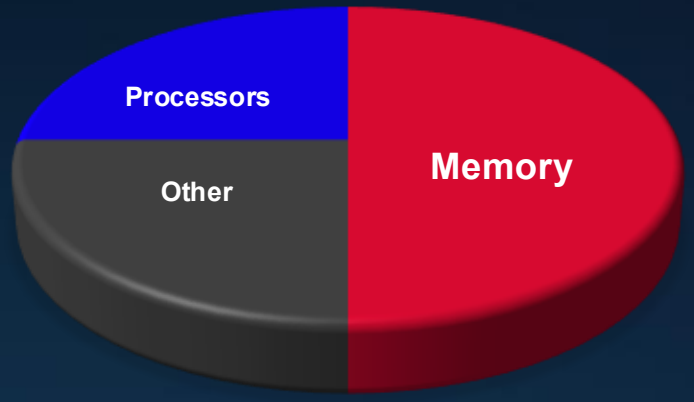
MEXT

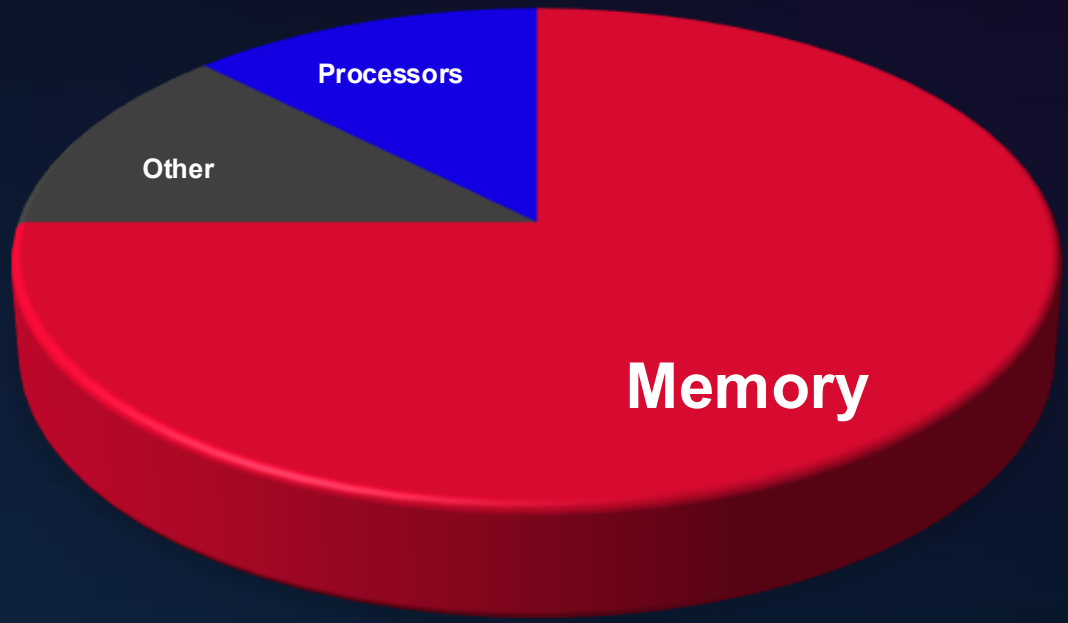
Smart Memory Tiering

Harnessing AI to radically transform the cost, capacity, and availability of memory

Gary Smerdon
Founder & CEO, MEXT
gary.smerdon@mext.ai

DRAM Has Become a Global Crisis

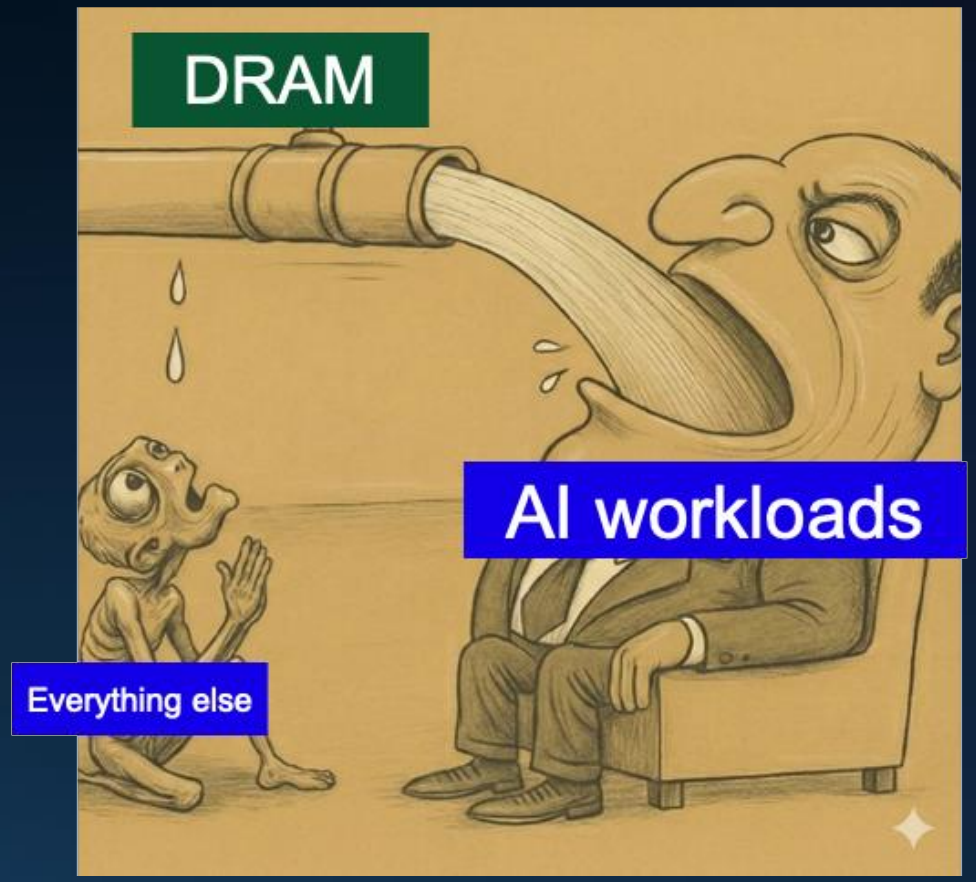




2023:
DRAM accounts for 50% of cost of each server

2026:
DRAM accounts for 60-90% of cost of each server

MEXT


DRAM
AI workloads
Everything else


RAM I WANT TO BUY
VENGEANCE
RAM I CAN BUY
RAM

MEXT

But Over Half of DRAM Is Wasted!

Azure

50%

Percentage of DRAM that
Half of VMs NEVER access

Azure

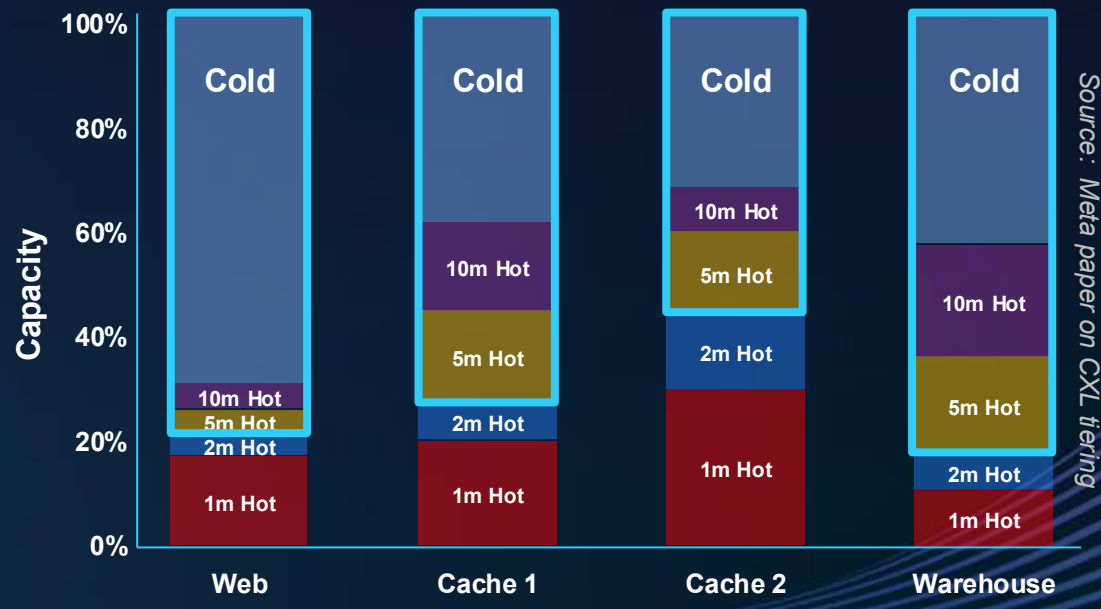
25%

Stranded memory as a percentage
of total DRAM

Meta

50-80%

Percentage of memory pages that go cold
(representative of many environments!)

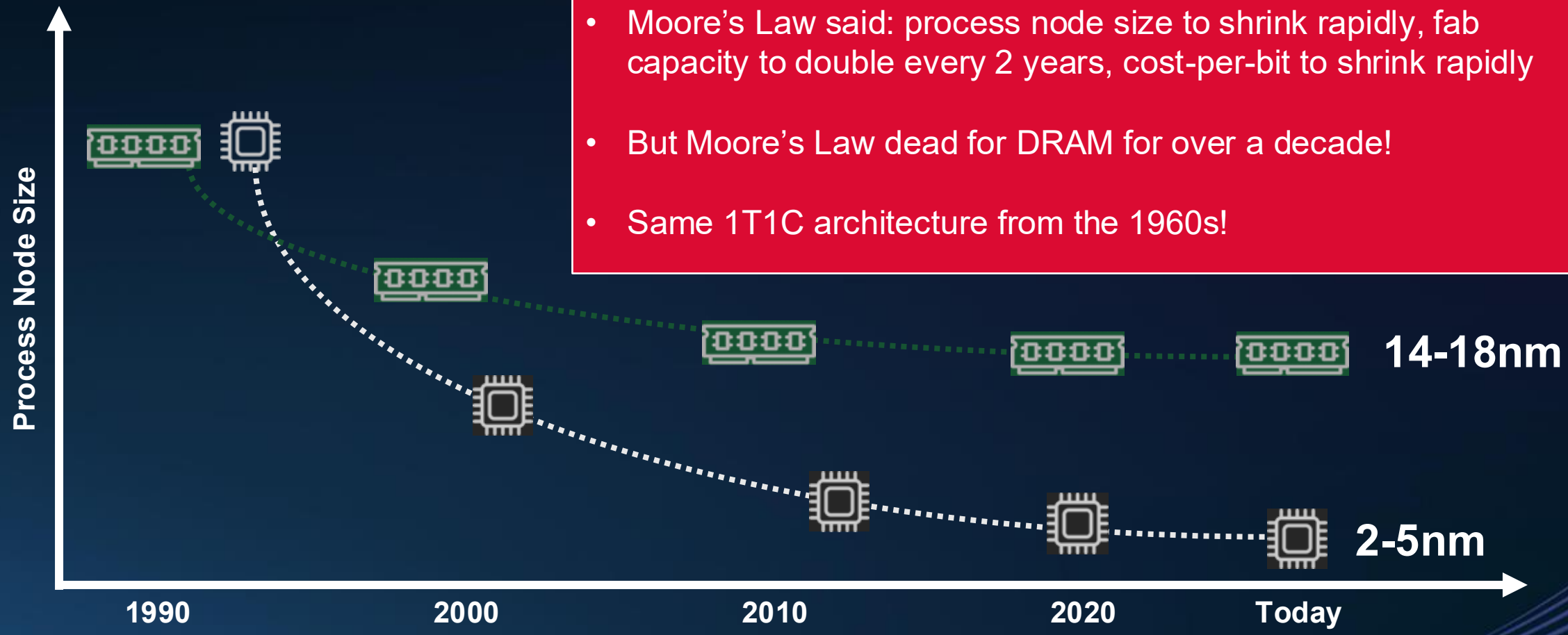


Source: Meta paper on CXL tiering

MEXT

Why The DRAM Problem Isn't Going Away

- Moore's Law said: process node size to shrink rapidly, fab capacity to double every 2 years, cost-per-bit to shrink rapidly
- But Moore's Law dead for DRAM for over a decade!
- Same 1T1C architecture from the 1960s!

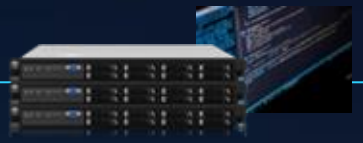


MEXT

How We Revolutionize Memory Cost and Capacity

Increase DRAM utilization

Eliminate overprovisioning and stranded DRAM



No hardware or software changes

HW / architectural and SW changes drive up cost

Bring flash into memory tier

Flash is 50X cheaper and 30X lower power vs DRAM

MEXT

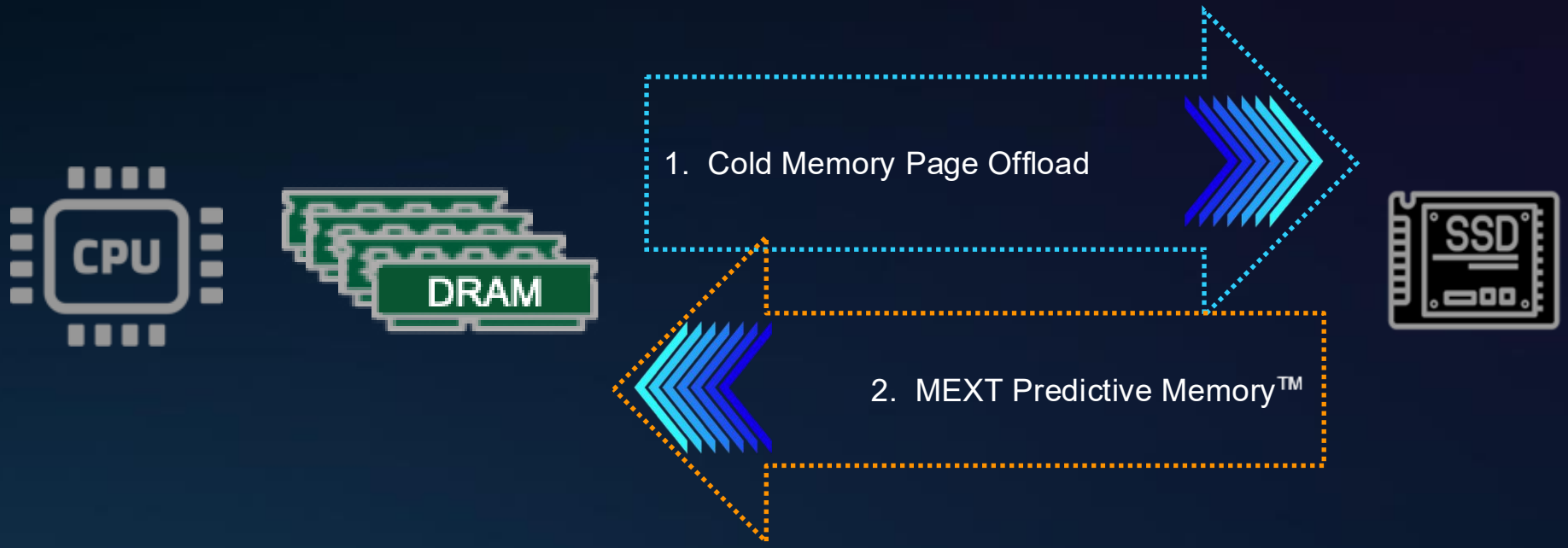
“Great idea, but it won’t work…”

Flash is 500X slower than DRAM

…until MEXT Predictive Memory™

MEXT

MEXT Predictively & Transparently Tiers Memory



1. MEXT transparently moves cold pages into NVMe flash
2. MEXT AI engine predicts and pushes pages back before they are needed

MEXT

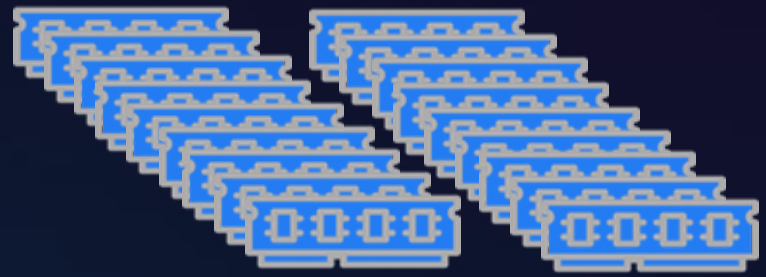
https://www.youtube.com/watch?v=zg-uBMsV_Yg

MEXT Makes Flash Appear as Memory

Constrained memory

MEXT
Predictive Memory™ Software

Expanded memory



MEXT Plug-In

- Lightweight, low-overhead Linux Driver

MEXT Predictive Memory™ Engine

- Ensemble of AI models enabling transparent auto-tiering of memory pages
- Continual training & inferencing using only a single CPU core

MEXT View™ Observability Tool

- Illustrates memory utilization and MEXT AI prediction accuracy using Open Telemetry (OTel)

MEXT

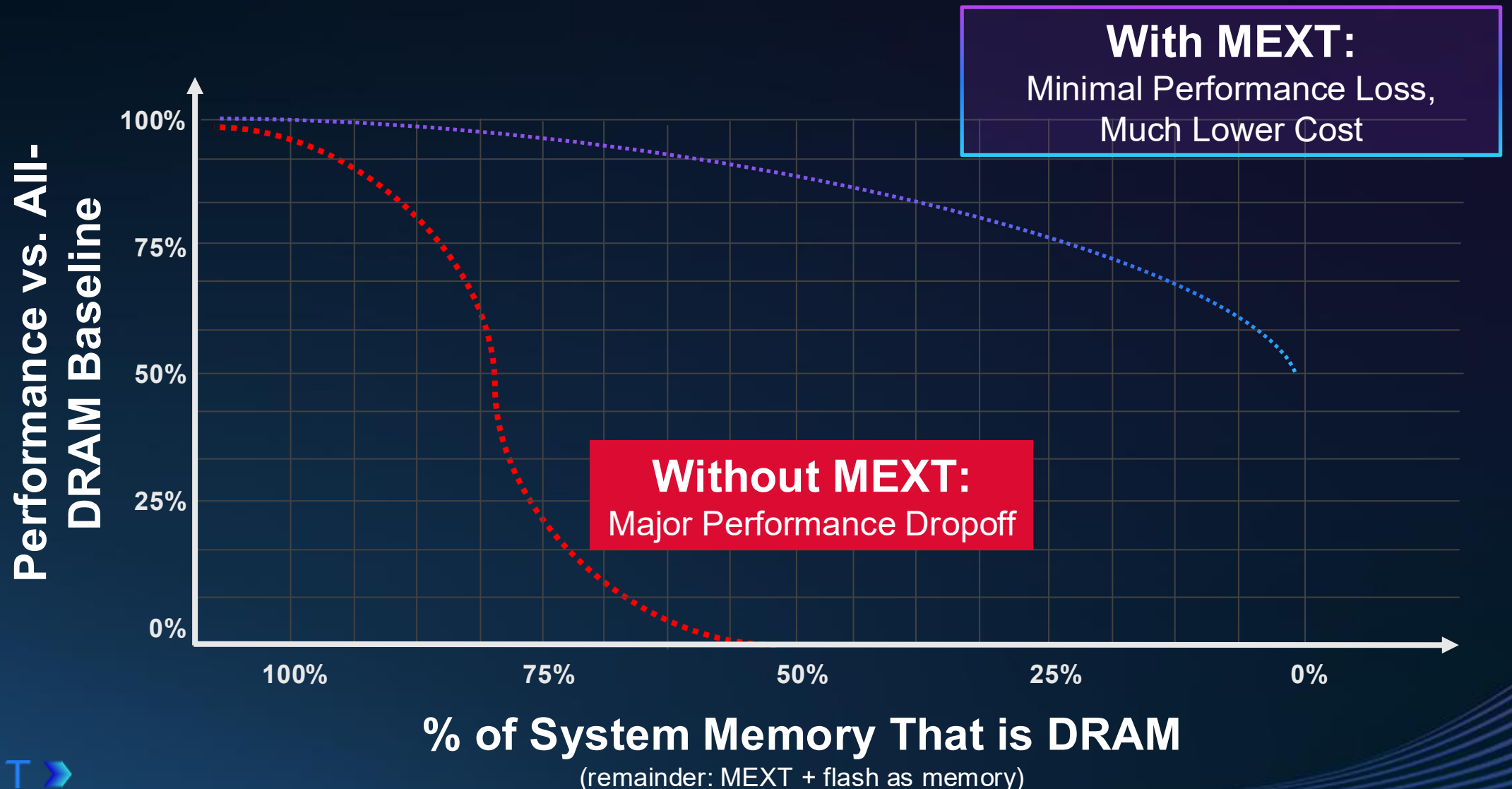
Revolutionizing Cost Savings On-Premises

Total Memory	Dell R6725 ALL DRAM	Dell R6725 DRAM + MEXT Memory™
1.5TB	$99K*	$30K* 70% Savings
3TB	No Longer Available	$68K Priceless
6TB	No Longer Available	$111K Priceless

MEXT

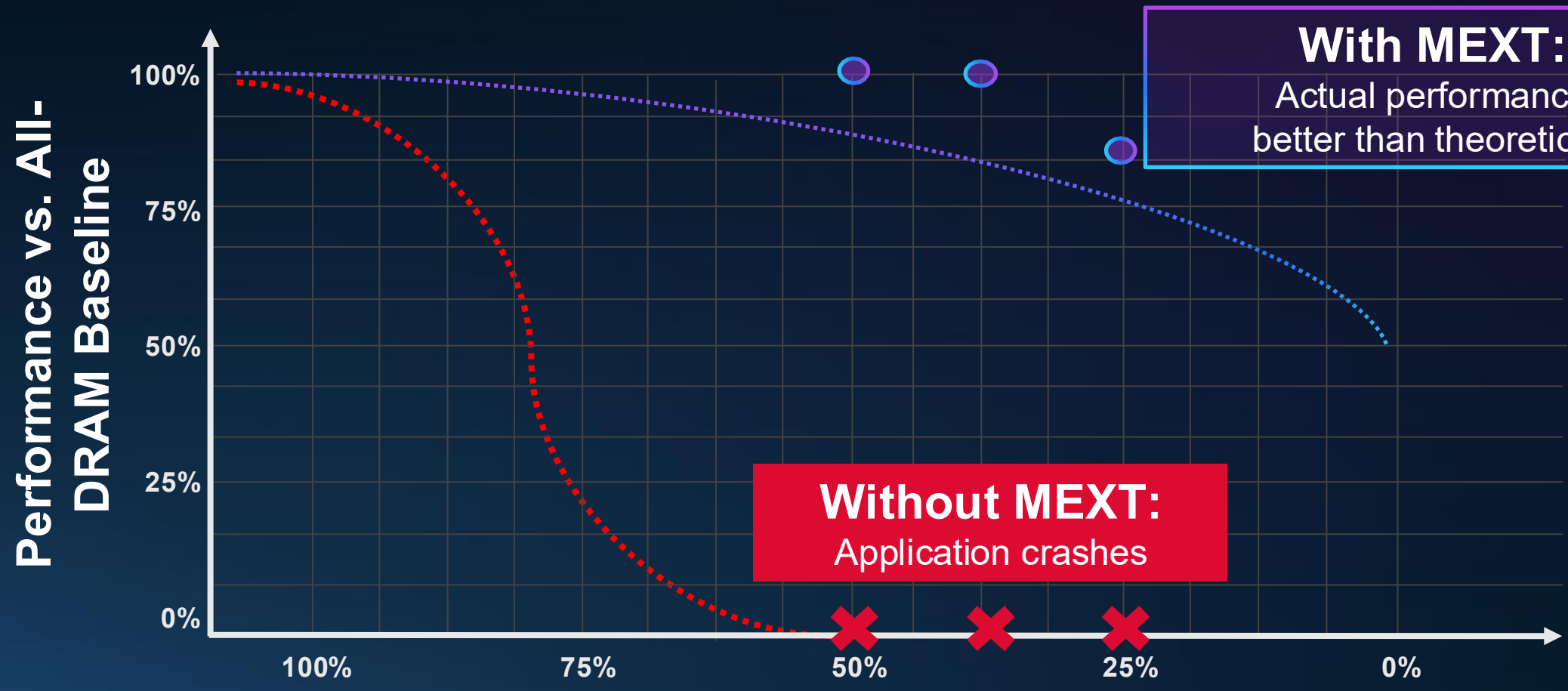
*Pricing as of Feb 1st Dell server config website

Theoretical Performance



MEXT

Actual Performance: Synopsys PrimeTime (EDA)



With MEXT:
Actual performance better than theoretical!

Without MEXT:
Application crashes

Hardware Config:
- AMD EPYC™ 9755 128-core processor
-12x 6300MT/s RAM
-KIOKIA NVMe drive

Software Stack:
-RHEL 8.10
-Linux Kernel 6.17.1
-MEXT Predictive Memory™
-Synopsys PrimeTime

Cadence Tempus: 2 Jobs, 1.8x Faster

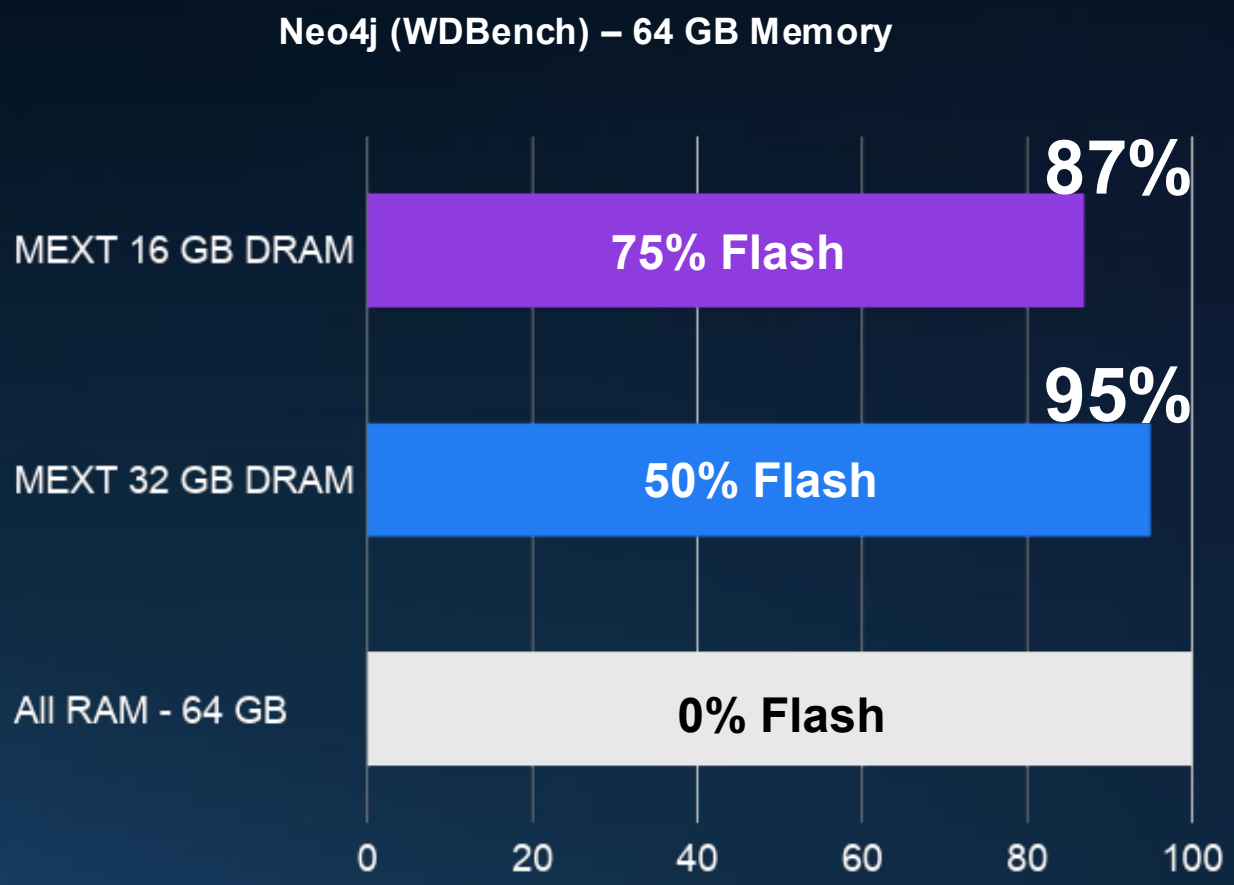
Same Hardware – 8 Hours Faster

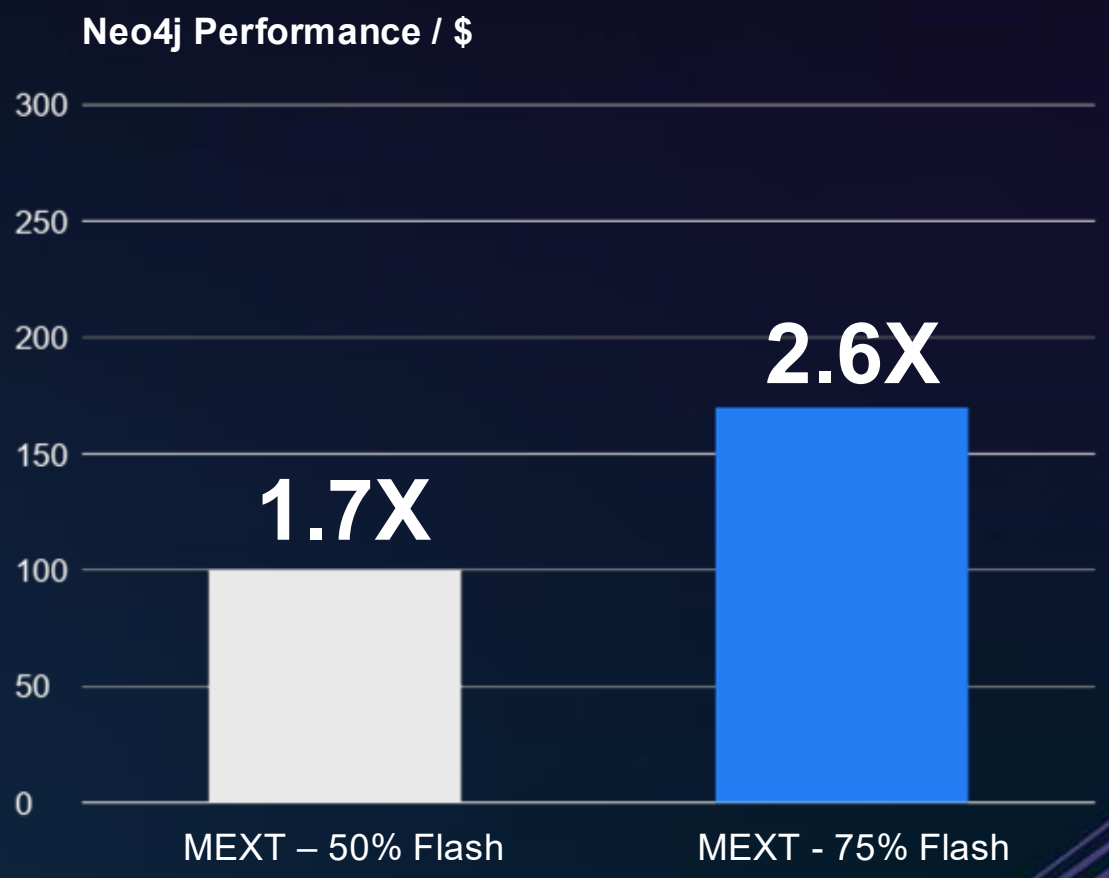
Hardware Config:
- 2x Intel Xeon Gold 6248 @ 2.5 GHz
- 40 cores
- 1 TB DRAM

Software Stack:
-RHEL 8.10
-Linux Kernel 4.18
-MEXT Predictive Memory™
- Cadence Tempus

MEXT

Improved Economics for Example: Neo4j





MEXT Workload Examples

Analytics

Relational In-Memory Database

ORACLE

In-Memory Key/Caching Value

redis

MEMCACHED

EDA

SYNOPSYS®

cadence

Graph Database

neo4j

AI-Native Vector Database

weaviate

AI Models

DeepSpeed

Animation Rendering

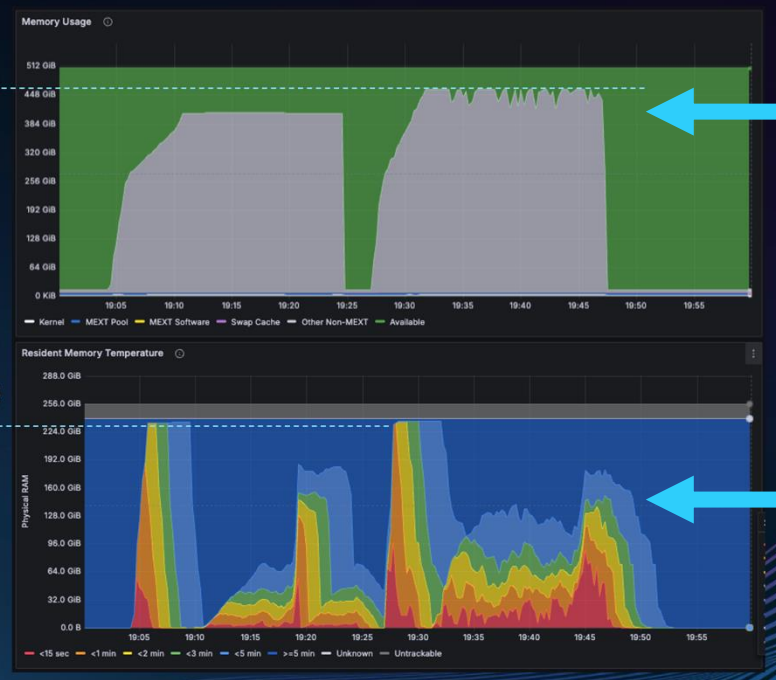
MOONRAY

SideFX®

MEXT View™: Actual customer results

Significant Memory Usage (~450GB)

Memory is very active (~256GB in <1min)



512GB Memory
256GB DRAM + 256GB MEXT

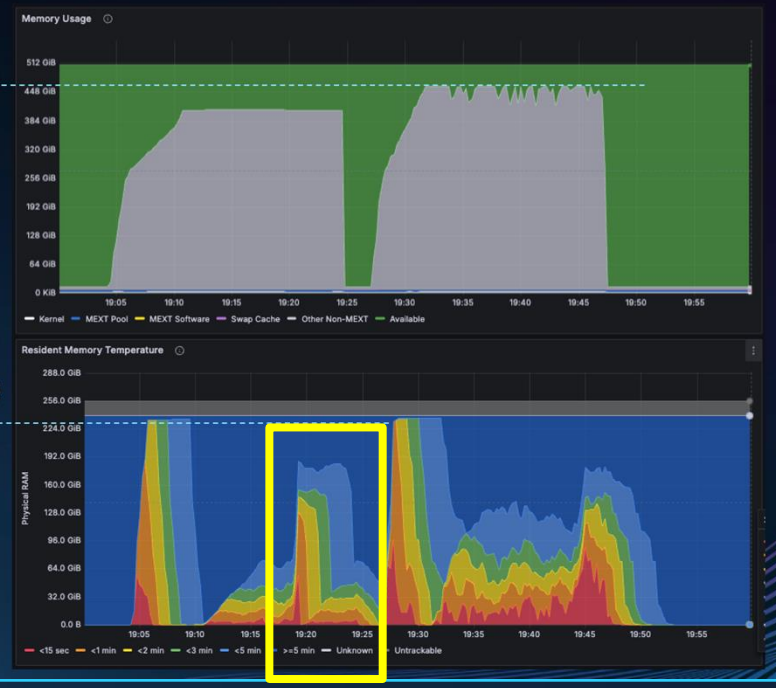
256GB DRAM
Most is not used for “long” periods

MEXT

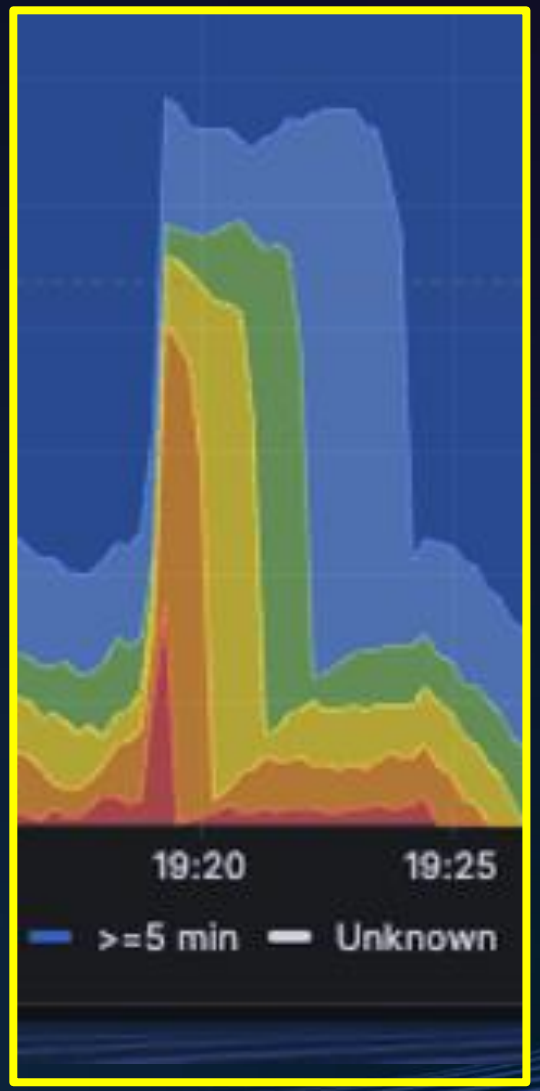
MEXT View™: Actual customer results

Significant Memory Usage (~450GB)

Memory is very active (~256GB in <1min)



MEXT



> 5 minute

< 5 minute

< 3 minute

< 2 minute

< 1 minute

<15 seconds

MEXT

In a Nutshell: MEXT Value

>2X Perf / $

Delivering more bang for the buck
on-prem or in the cloud

Abundant Memory

“Software expands to fill available memory” (Wirth’s Law),
avoid sharding of large datasets for better performance

Lower Power

Basis in 30X lower-power
form factor (flash) vs DRAM

Server Life

More capacity on existing
servers, extending lifespan

Fast Deployment

Software-only, drops in,
any infra / configuration

Simplified SKUs

No lock-in to standard
memory sizes, scale easily

MEXT

Fast Forward!

Gary Smerdon
Founder & CEO, MEXT
gary.smerdon@mext.ai

MEXT