©2023 Flash Memory Summit. All Rights Reserved

Flash Memory Summit

Analog eFlash Drives AI/ML Acceleration

AIML-102-1

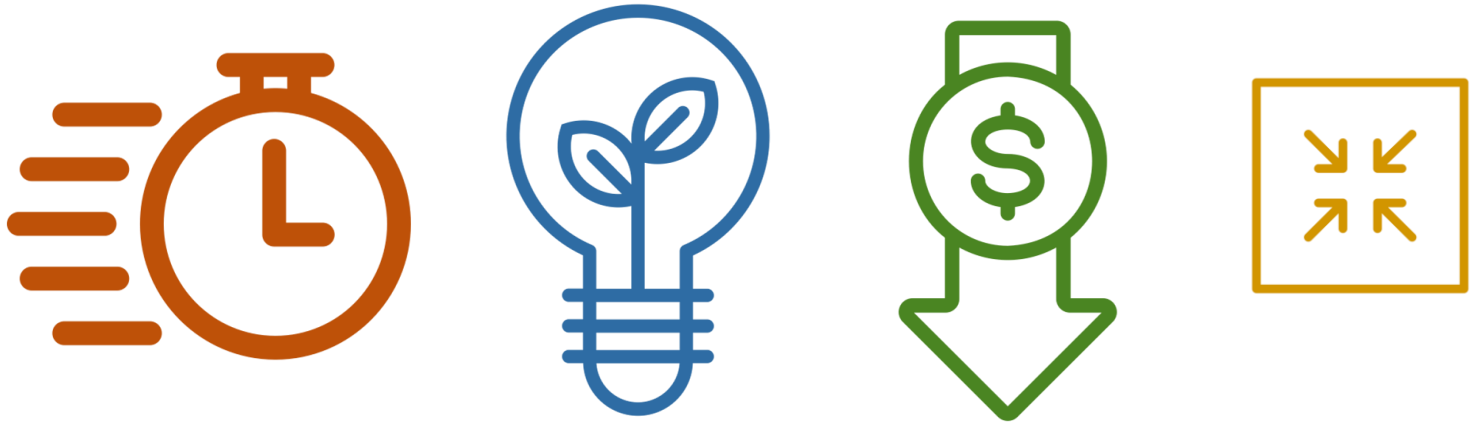
Presenter: Dave Eggleston Microchip/SST

What's So Challenging About Digital Edge Inference?

Flash Memory Summit

Want:

- Fast
- Power efficient
- Low cost
- Small



Get (1 or more):

- Slow
- Power Hog
- Expensive
- Big

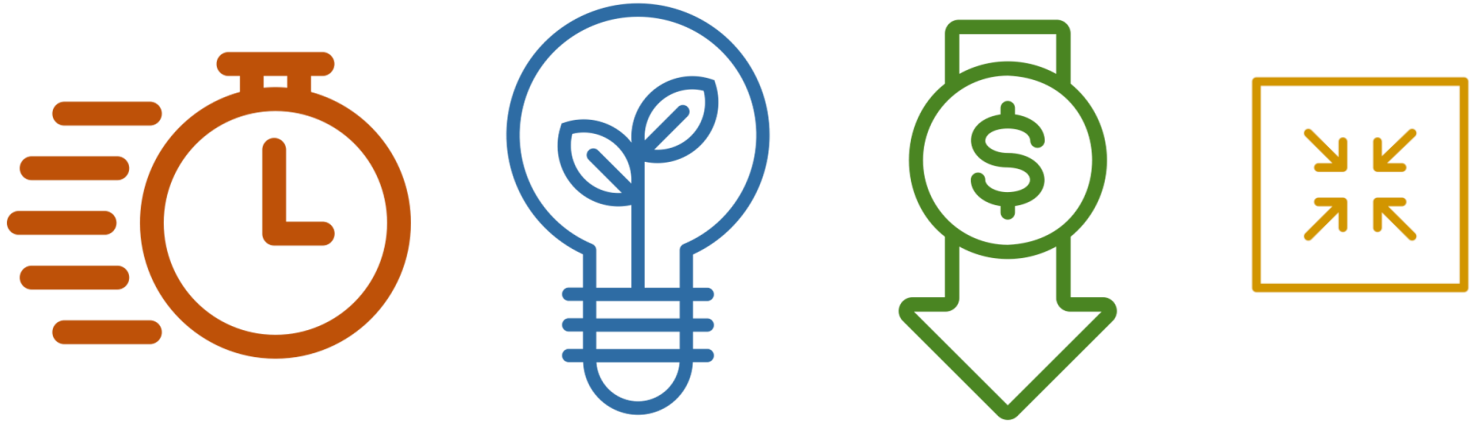
©2023 Flash Memory Summit. All Rights Reserved

What's So Challenging About Digital Edge Inference?

Flash Memory Summit

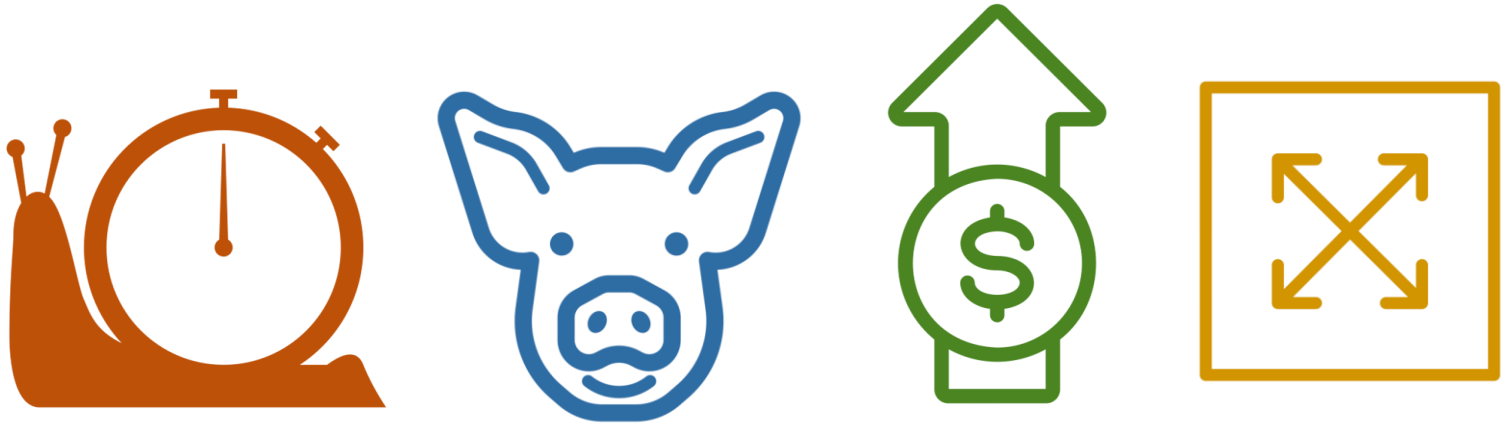
Want:

- Fast
- Power efficient
- Low cost
- Small



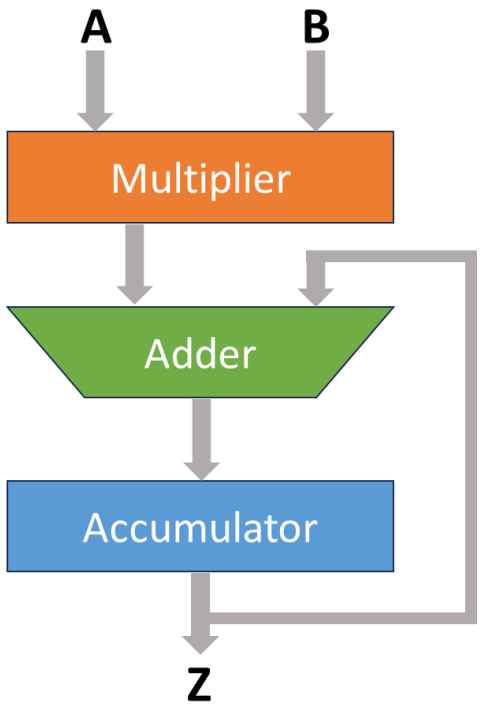
Get (one or more):

- Slow
- Power hog
- Expensive
- Big



©2023 Flash Memory Summit. All Rights Reserved

Why? Multiply and Accumulate (MAC) Operation Analysis



Read A
Read B
Multiply
Read Z
Add Z
Update (**write**) Z
Repeat…
Repeat…
Repeat…
…

Three memory **reads** and one memory **write** every loop

AlexNet → **3B** DRAM accesses!

Digital MAC

Write each array element with conductance **G**

Input V
Output I
DONE!
Next layer…

AnyNet → **Zero** DRAM accesses!

Single cycle MAC operations!

Analog MAC

Source: "Multiply accumulate operations in memristor crossbar arrays for analog computing", Journal of Semiconductors (2021), Chen, et al

©2023 Flash Memory Summit. All Rights Reserved

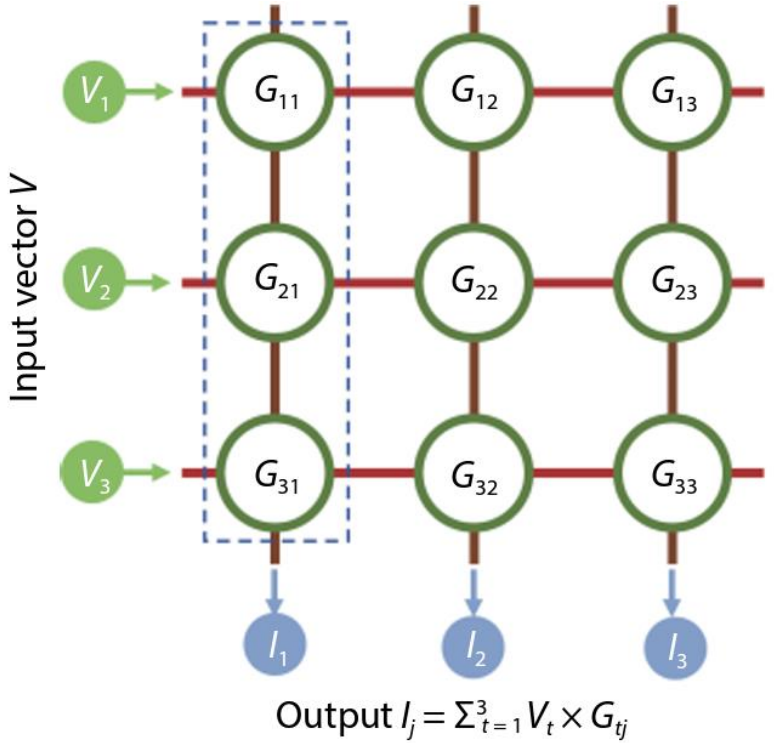
Why? Multiply and Accumulate (MAC) Operation Comparison

Read A
Read B
Multiply
Read Z
Add Z
Update (**Write**) Z
Repeat…
Repeat…
Repeat…
…

AlexNet → **3B** DRAM accesses!

3 memory **Reads** & 1 memory **Write** every loop

Digital MAC



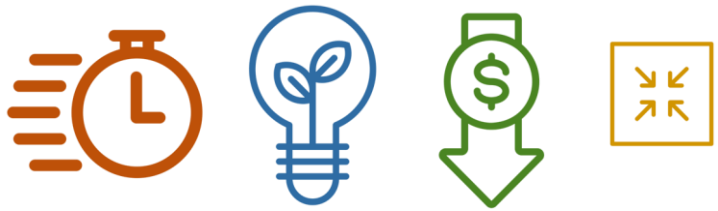
Output $I_j = \Sigma_{t=1}^{3} V_t \times G_{tj}$

AnyNet → **Zero** DRAM accesses!

Write each array element with conductance **G**

Input V
Output I
DONE!
Next layer…

Single cycle MAC operations!

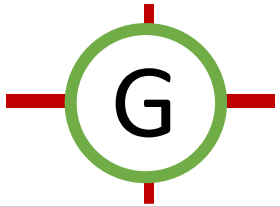


Analog MAC

Source: "Multiply accumulate operations in memristor crossbar arrays for analog computing", Journal of Semiconductors (2021), Chen, et al

©2023 Flash Memory Summit. All Rights Reserved

eFlash is the *BEST* Choice for Analog MAC Array Element



Flash Memory Summit

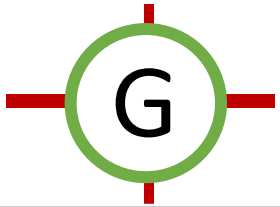
	eFlash	eRRAM	eMRAM	ePCM	FeFET
Bits per cell capability	5 bpc+	2bpc	1bpc	2bpc	TBD
Foundry nodes available	>100	1	5	0	0
Production devices shipped	100B+	0+	1M?	N/A	N/A
Analog capable	YES	Maybe	No		
Analog IP available	YES	No	N/A		
Analog IP demos	YES	N/A			
Analog IP in production	YES				

SST

memBrain™

©2023 Flash Memory Summit. All Rights Reserved

eFlash is the *BEST* Choice for Analog MAC Array Element



Flash Memory Summit

	eFlash	eRRAM	eMRAM	ePCM	FeFET
Bits per cell capability	5 bpc+	2 bpc	1 bpc	2 bpc	TBD
Foundry nodes available	>100	1	5	0	0
Production devices shipped	100B+	0+	1M?	N/A	N/A
Analog capable	YES	Maybe	No		
Analog IP available	YES	No	N/A		
Analog IP demos	YES	N/A			
Analog IP in production	YES				

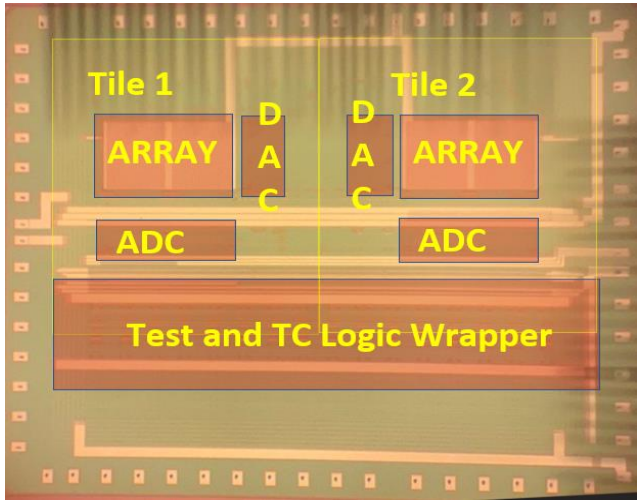
SST

memBrain™

©2023 Flash Memory Summit. All Rights Reserved

eFlash 5 bpc (PLC) Demonstrated in 2X nm Silicon

- Silicon fully functional (2021)
  - All macro functions validated
  - 5-bits per cell = 32 levels
  - Read neural multiply operation realized
  - Neural net functionality realized over temp – MNIST ~95% (MLP), ~97% (CNN)



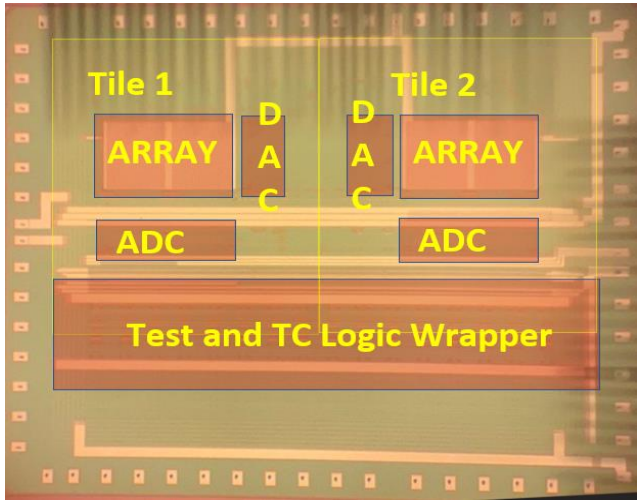
- 5 bpc MW and bit placement
  - 0 to 100 nA range
  - Sigma <1.3%
    - Single electron resolution

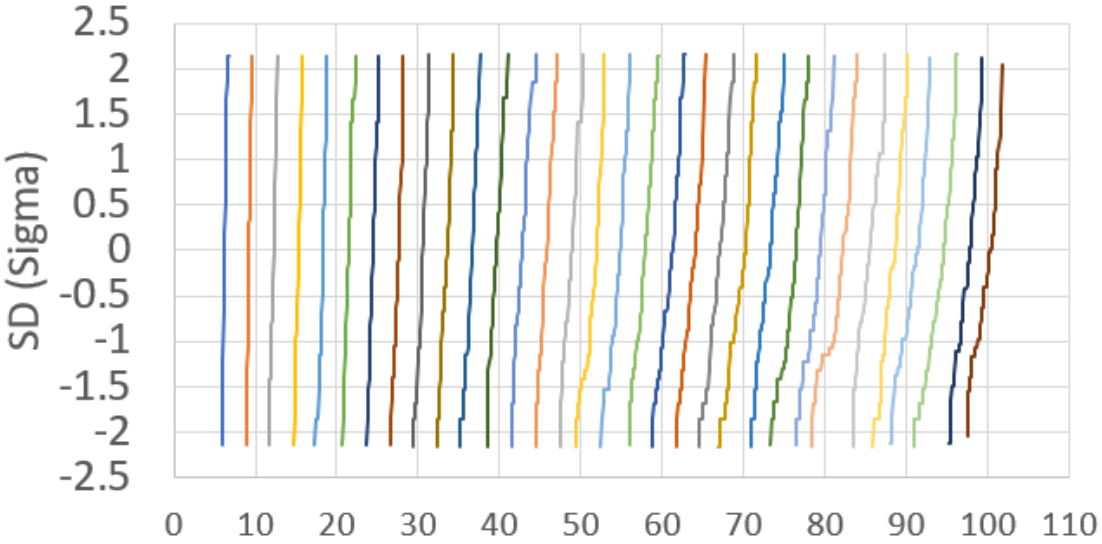
©2023 Flash Memory Summit. All Rights Reserved

Source: "*Computing-in-Memory with SuperFlash® memBrain™ Technology*," Symp. VLSI 2022; N. Do, M. Reiten, H. Tran; Microchip/SST

eFlash 5 bpc (PLC) Demonstrated in 2X nm Silicon

- Silicon fully functional (2021)
  - All macro functions validated
  - 5-bits per cell = 32 levels
  - Read neural multiply operation realized
  - Neural net functionality realized over temp – MNIST ~95% (MLP), ~97% (CNN)
- 5 bpc MW and bit placement
  - 0 to 100 nA range
  - Sigma <1.3%
    - Single electron resolution





Source: "*Computing-in-Memory with SuperFlash® memBrain™ Technology*," Symp. VLSI 2022; N. Do, M. Reiten, H. Tran; Microchip/SST

©2023 Flash Memory Summit. All Rights Reserved

Analog MACs Handle Large Weights with Ease!

- Why is using analog superior for edge inference acceleration?
  - Parallel in-memory compute the matrix vector product is very power efficient (no data movement)
  - Analog matrix vector product is computed in constant time
  - Large weight matrix takes the same time to compute the matrix vector product as a small weight

- Strategy divide and conquer!
  - Small tensor input; use Digital
  - Large tensor input; use Analog

Source: IBM, https://aihwkit.readthedocs.io/en/latest/pros_cons.html

©2023 Flash Memory Summit. All Rights Reserved

Analog MACs Handle Large Weights with Ease!

- Why is using analog superior for edge inference acceleration?
  - Parallel in-memory compute the matrix vector product is very power efficient (no data movement)
  - Analog matrix vector product is computed in constant time
  - Large weight matrix takes the same time to compute the matrix vector product as a small weight

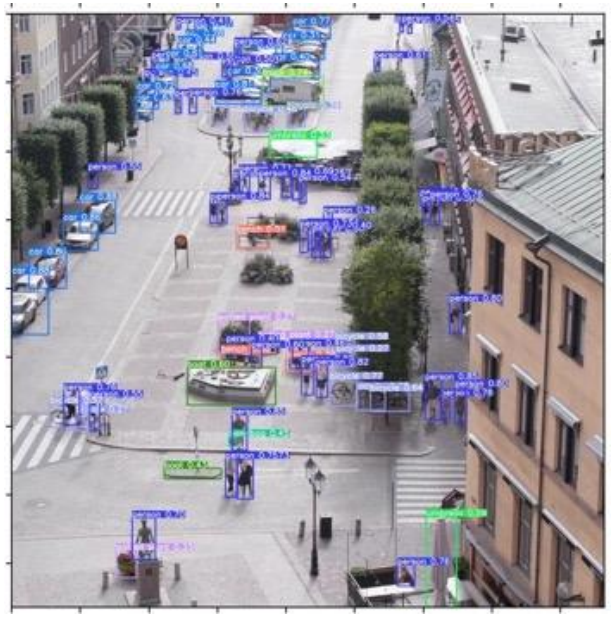
- Strategy divide and conquer!
  - Small tensor input; use Digital
  - Large tensor input; use Analog

©2023 Flash Memory Summit. All Rights Reserved

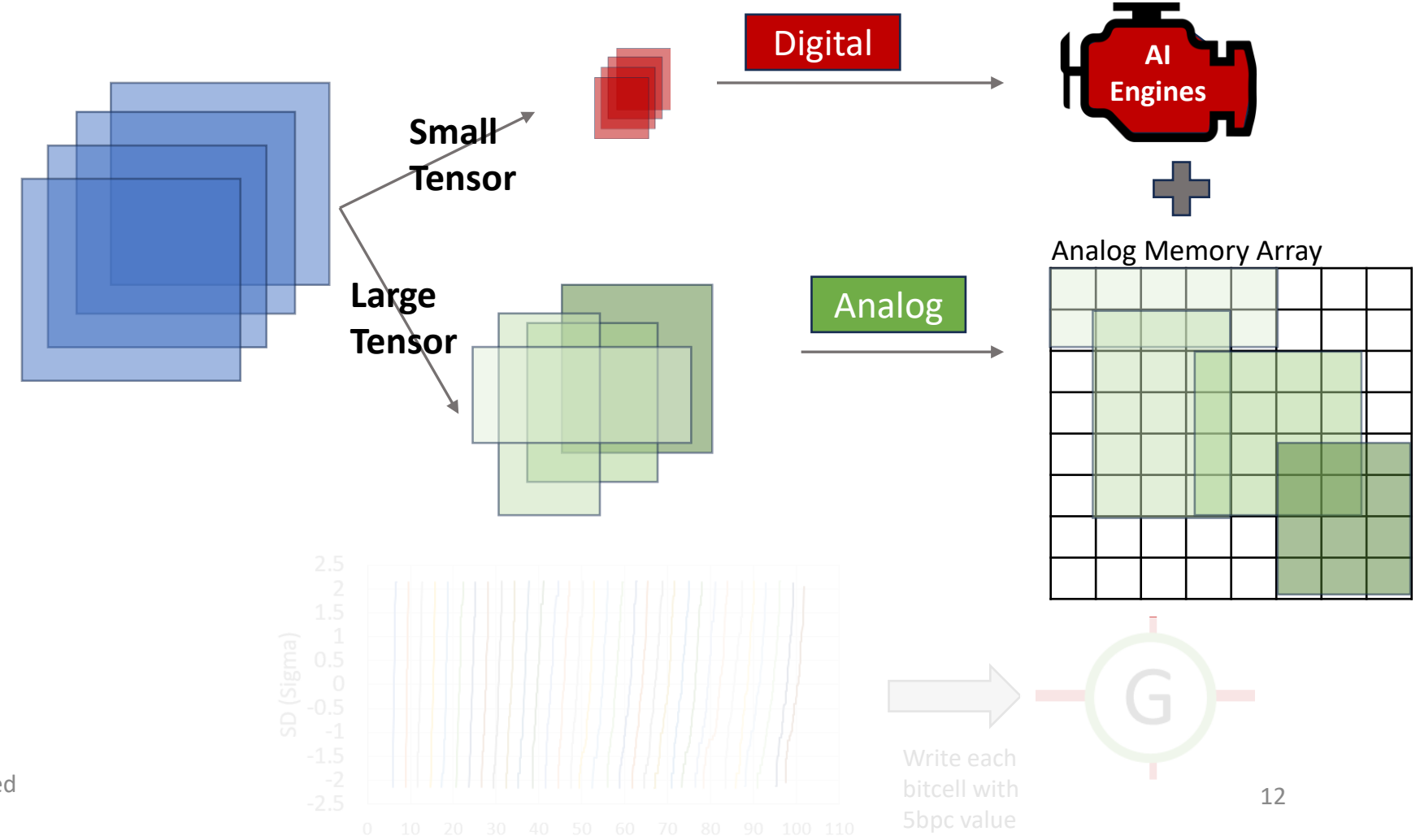
Source: IBM, https://aihwkit.readthedocs.io/en/latest/pros_cons.html

Divide the Tensor Inputs Between Analog and Digital

- Example: Video object detection YOLOv5



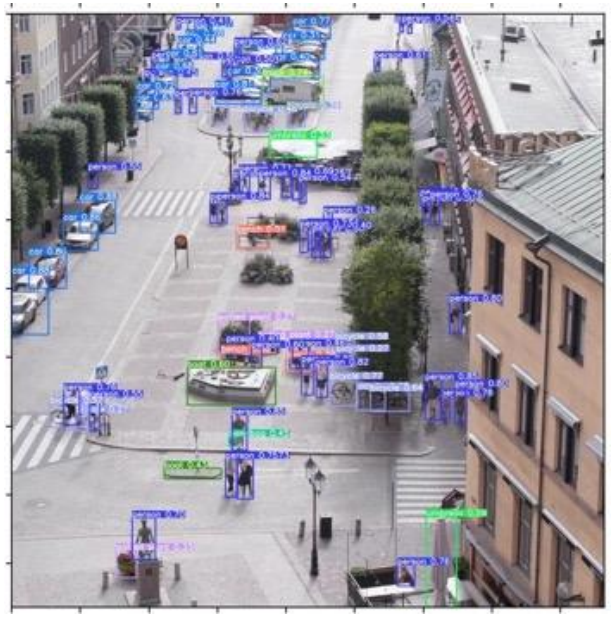
Security video outside a train station



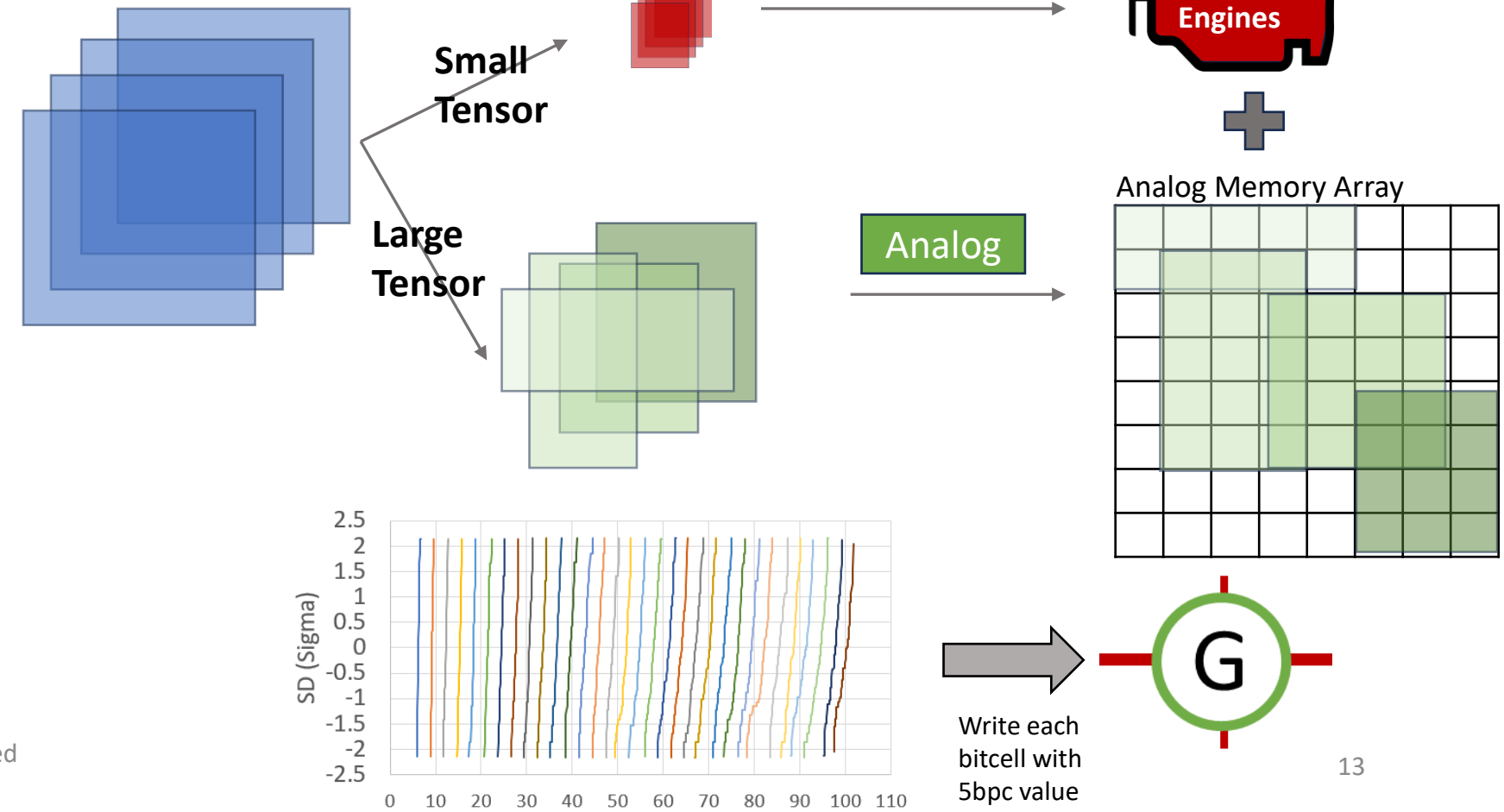
©2023 Flash Memory Summit. All Rights Reserved

Divide the Tensor Inputs Between Analog and Digital

- Example: Video object detection YOLOv5



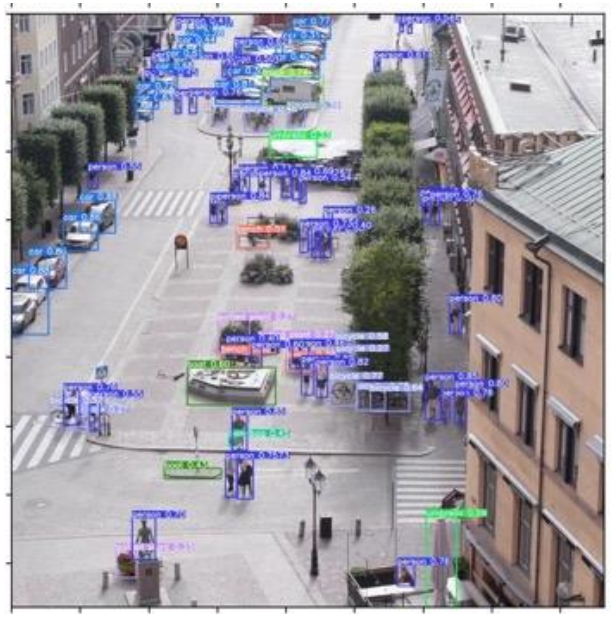
Security video outside a train station



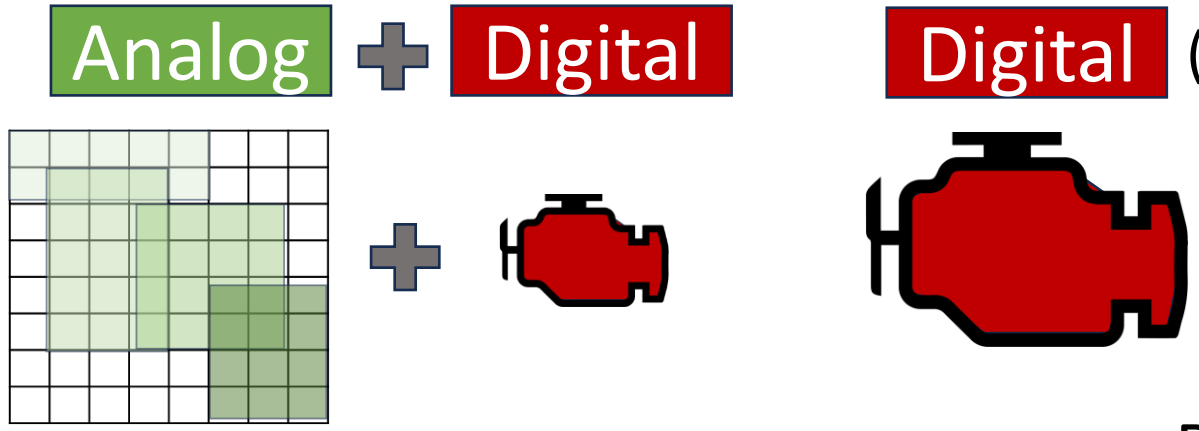
©2023 Flash Memory Summit. All Rights Reserved

Performance Difference: Analog vs. Digital

- Example: Video object detection YOLOv5



Security video outside a train station



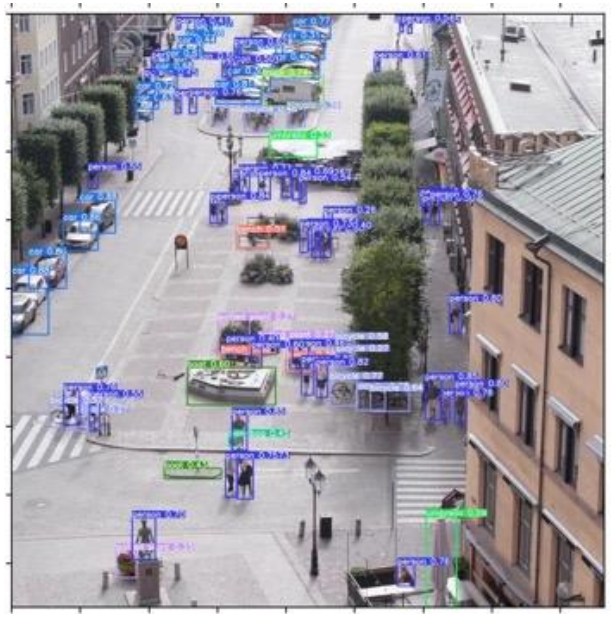
			Difference
Die Cost	1/32	1	97%↓
Frames per sec	150 fps	200 fps	75%↓
Power	3W	60W	20x↓
FPS/W	50	3.33	15x↑
FPS/$	3	0.125	24x↑

©2023 Flash Memory Summit. All Rights Reserved

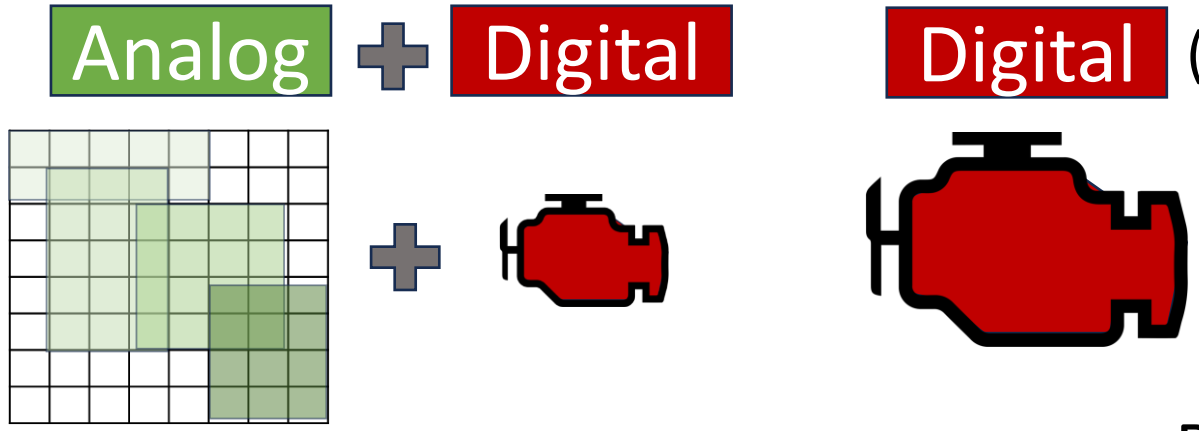
Performance data courtesy of Mythic Inc. an SST memBrain licensee.

Performance Difference: Analog vs. Digital

- Example: Video object detection YOLOv5



Security video outside a train station

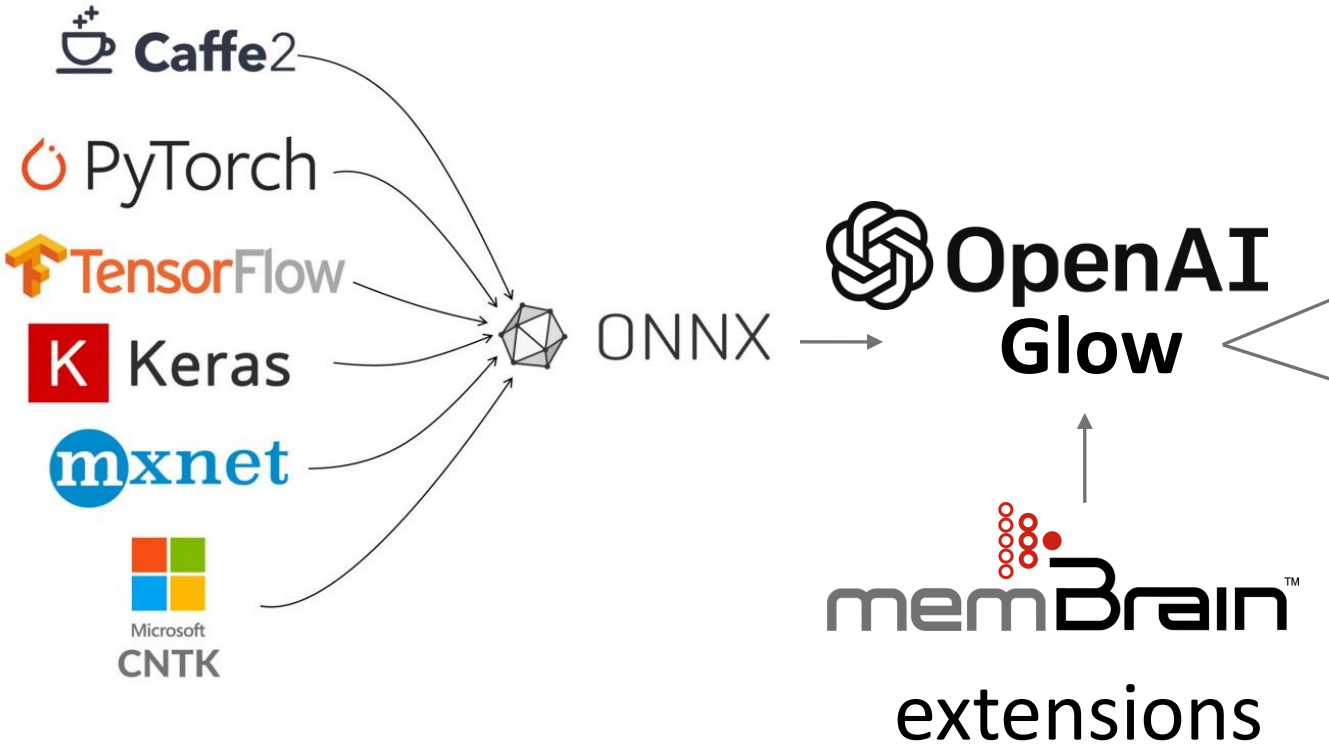


			Difference
Die Cost	1/32	1	97%↓
Frames per sec	150 fps	200 fps	75%↓
Power	3W	60W	20x↓
FPS/W	50	3.33	15x↑
FPS/$	3	0.125	24x↑

©2023 Flash Memory Summit. All Rights Reserved

Performance data courtesy of Mythic Inc. an SST memBrain licensee.

Software Flow Example: eFlash Analog Hardware Aware



Training Outside SoC

Mapping Inside SoC for Inference

©2023 Flash Memory Summit. All Rights Reserved

© Copyrights, trademarks, and logos owned by their respective companies

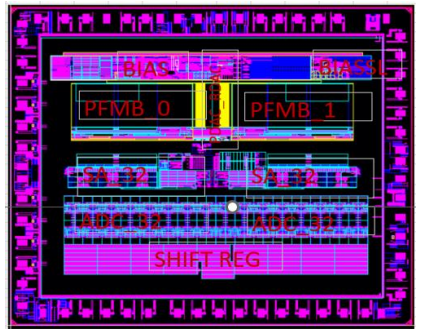
Software Flow Example: eFlash Analog Hardware Aware

Flash Memory Summit

Caffe2
PyTorch
TensorFlow
Keras
mxnet
Microsoft CNTK
ONNX
OpenAI Glow
memBrain™
extensions

Training Outside SoC

arm
Digital
memBrain™
Analog



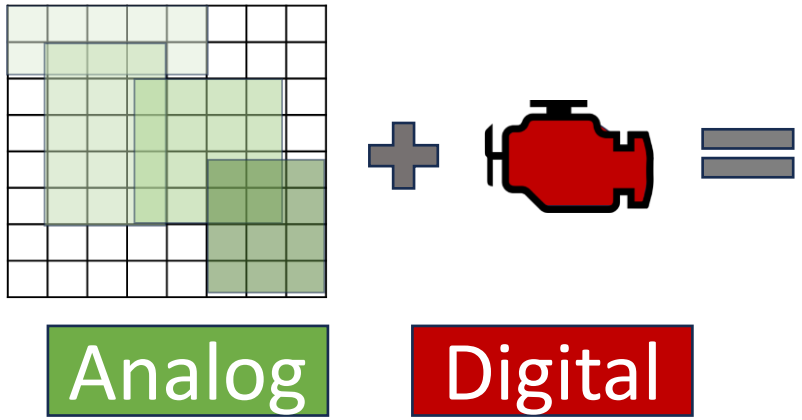
drivers

Mapping Inside SoC for Inference

©2023 Flash Memory Summit. All Rights Reserved

© Copyrights, trademarks, and logos owned by their respective companies

Best of Analog + Best of Digital = Best Edge Inference Solution!



Get:

- Fast
- Power efficient
- Low cost
- Small

©2023 Flash Memory Summit. All Rights Reserved

Best of Analog + Best of Digital = Best Edge Inference Solution!

Flash Memory Summit

Get:

- Fast
- Power efficient
- Low cost
- Small

SST®

memBrain™

©2023 Flash Memory Summit. All Rights Reserved

Want to see and learn more?

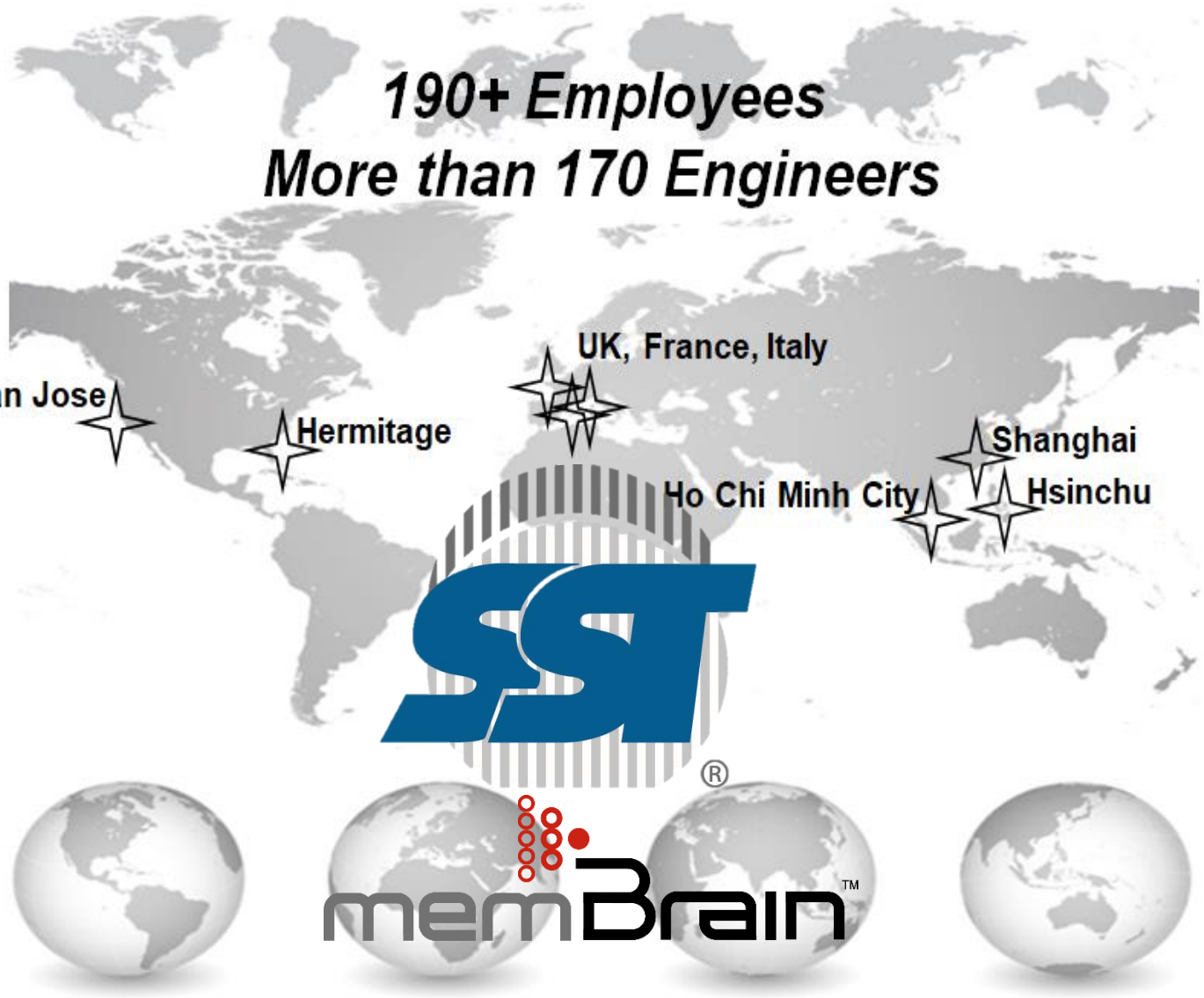
- Visit our SST demo booth at the AI/ML Edge Hardware Summit
- September 12 – 14 in Santa Clara
- aihwedgesummit.com

AI HARDWARE & EDGE AI SUMMIT

©2023 Flash Memory Summit. All Rights Reserved

© Copyrights, trademarks, and imaged owned by Kisaco Research Inc.

Thank You!



SST US/Japan Sales:
DEggleston@SST.com

SST EU/Korea Sales:
CBrown@SST.com

SST China/TW Sales:
GChen@SST.com

©2023 Flash Memory Summit. All Rights Reserved