STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

In-SRAM Computing For Lower Power LLMs

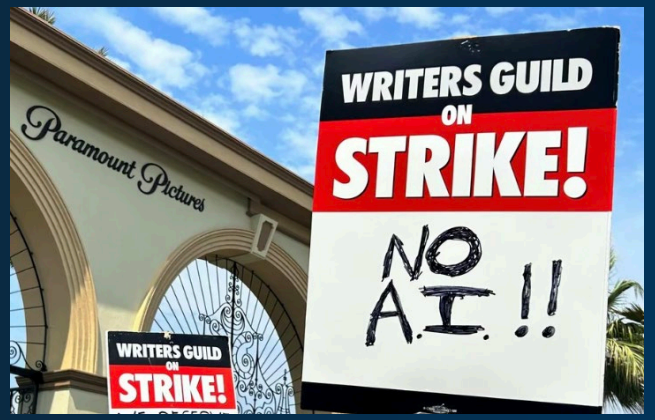
GSI Technology

George Williams, Head of Embedded AI

Generative AI In The News

Programmers, beware: ChatGPT has ruined your magic trick

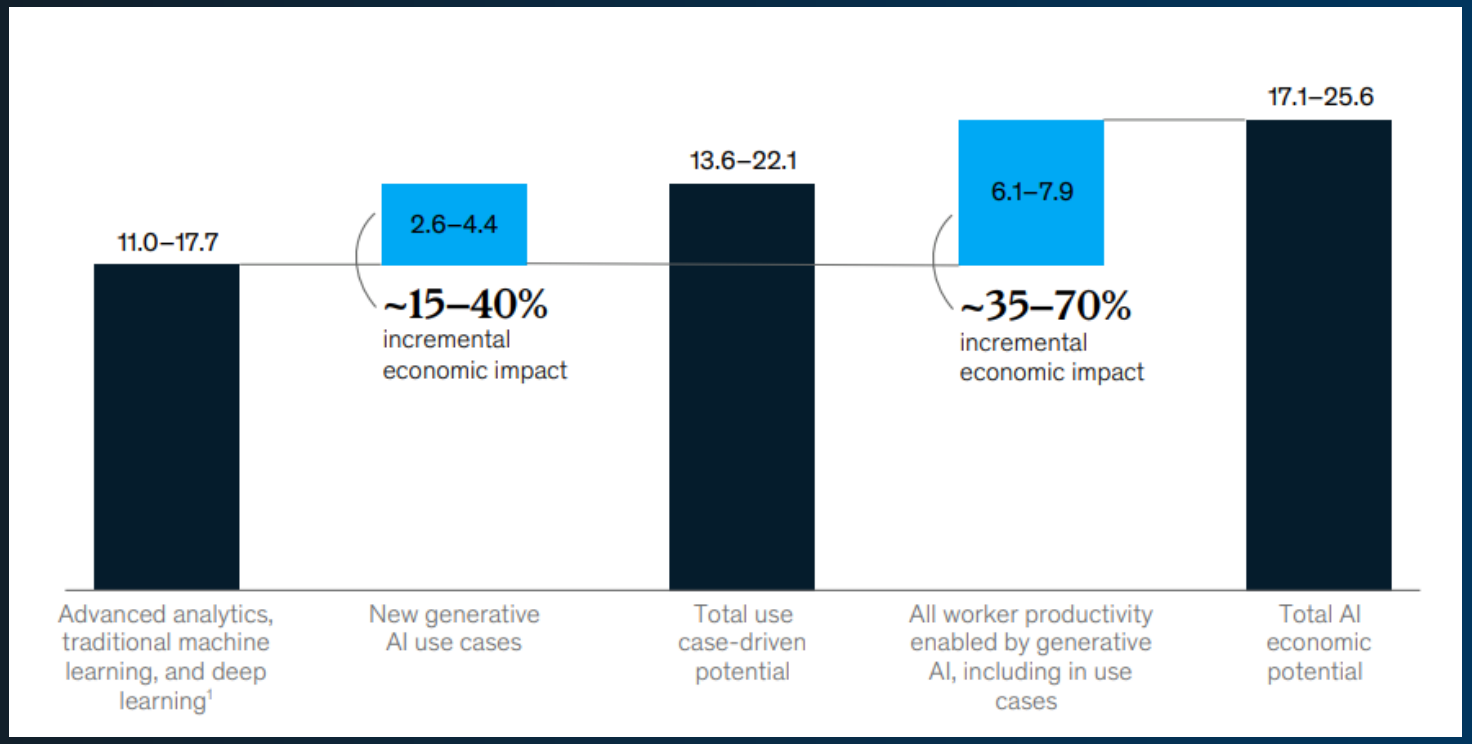
Generative AI Drugs Are Coming



It Was The Best Of Times, It Was The Worst Times…

Generative AI Impact

Mckinsey Co, 2023

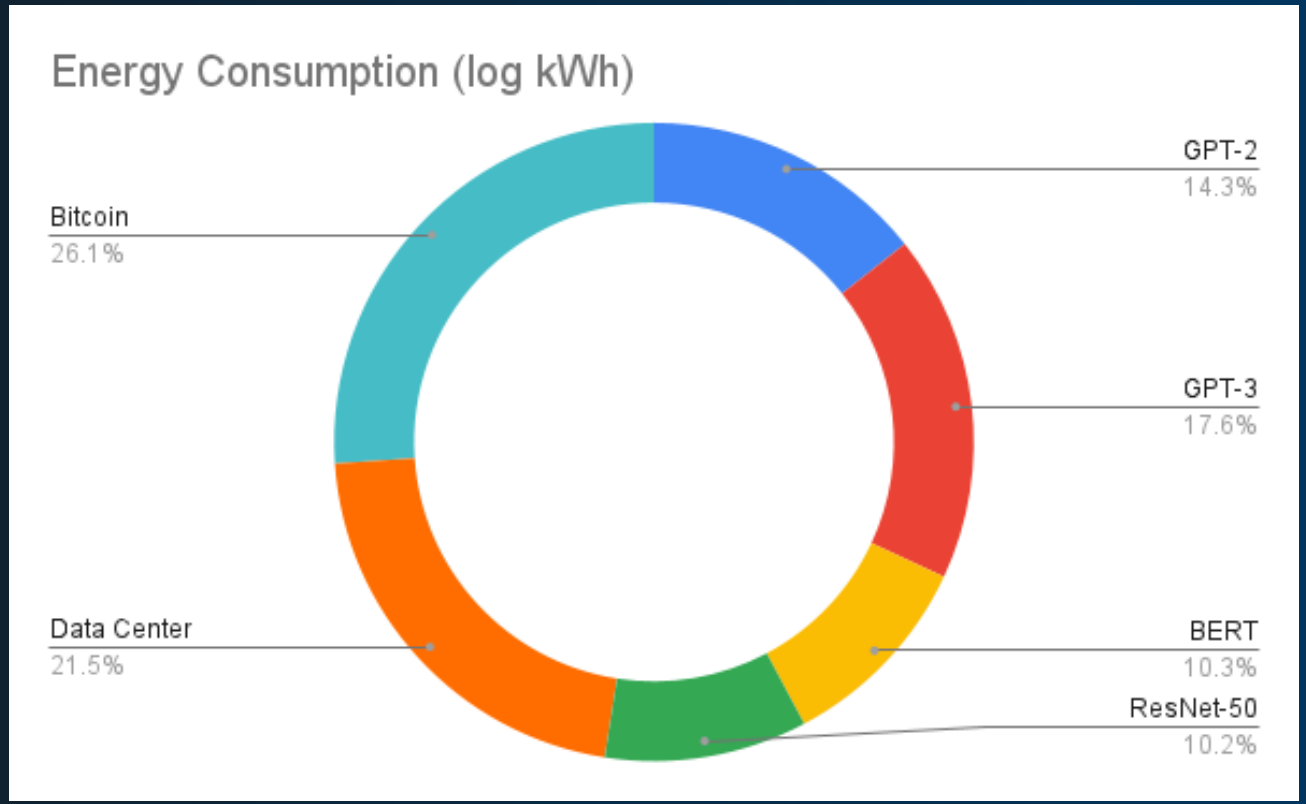


It Was The Best Of Times...

Energy Costs of Advanced Computing

https://www.nnlabs.org/power-requirements-of-large-language-models

Application	Energy Consumption
GPT-2	28,000 kWh
GPT-3	284,000 kWh
BERT	1,536 kWh
ResNet-50	1,500 kWh
Data Center	4,500 tons CO2
Bitcoin	121.36 TWh/year



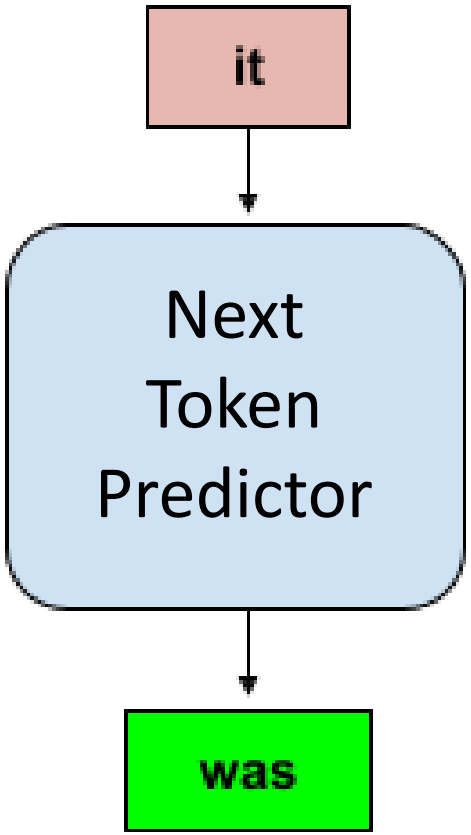
...It Was The Worst Of Times.

Agenda

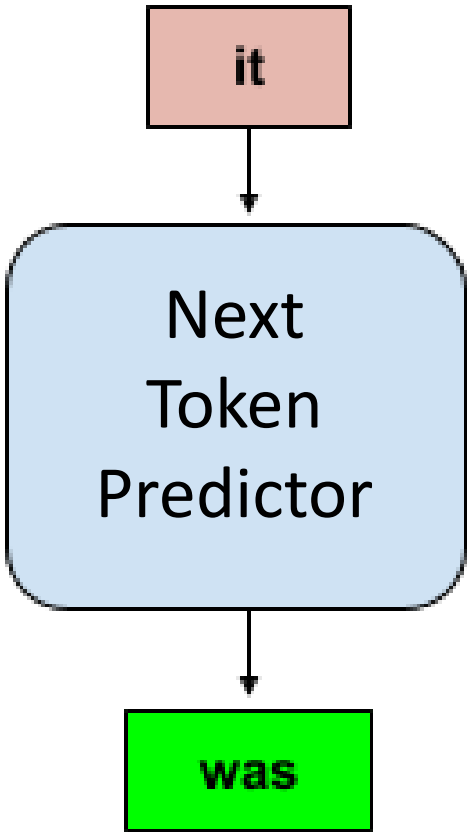
- Next Word Prediction
- Transformer Essentials
- Von-Neumann Architecture & Bottleneck
- New Paradigm: Adding Compute Into SRAM
- Associative Compute Grid Power
- Modular IP For Size and Power Budgets
- Token Rates
- Try It Out!!

Next Word Prediction

Neural Language Modeling

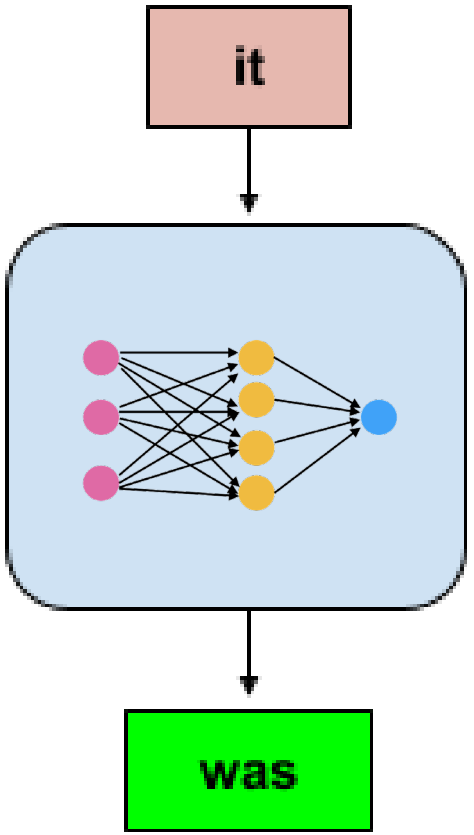


Neural Language Modeling



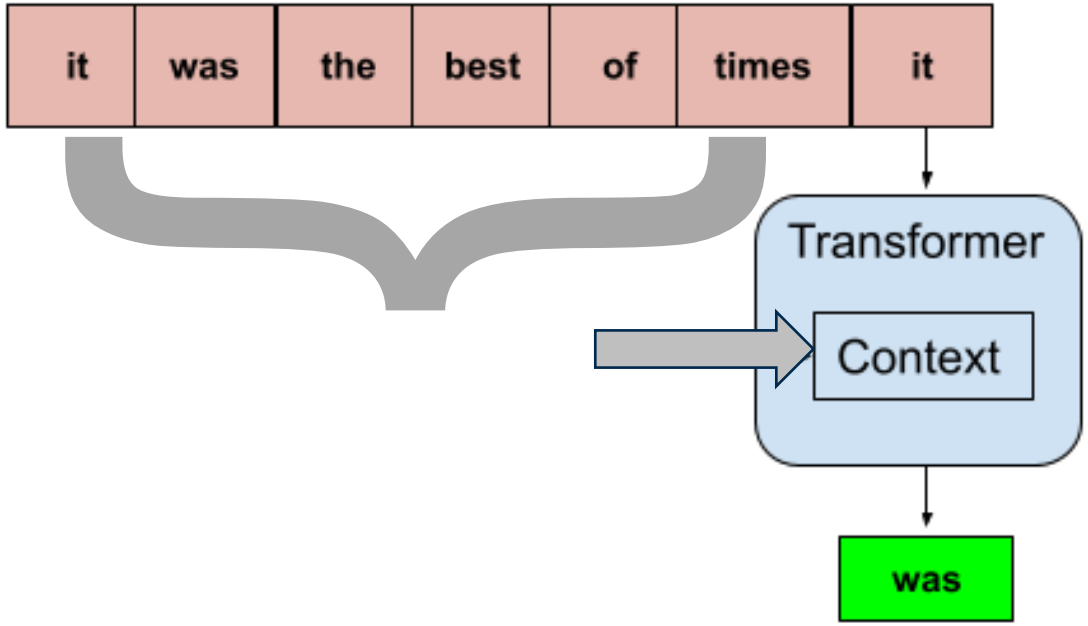
- task: *next token prediction*
- idea dates back to 70s

Neural Language Modeling



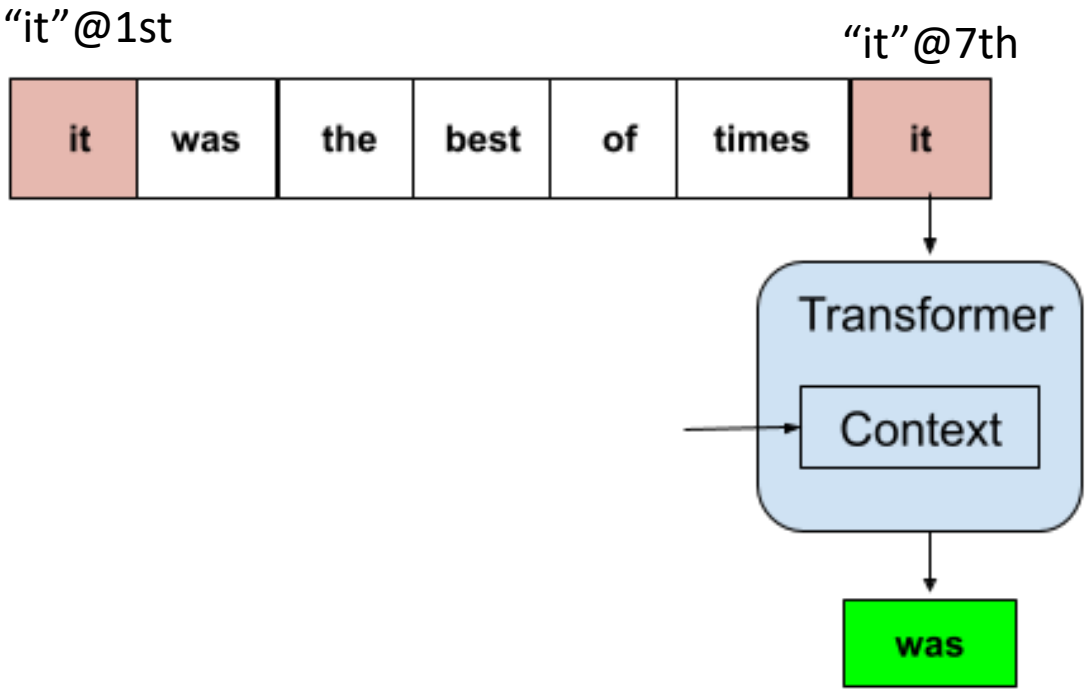
- task: *next token prediction*
- idea dates back to 70s
- 90s: RNNs, LSTMs, GRUs…
- nothing works well until ***Transformer***
- *wait…just next token?*

Neural Language Modeling



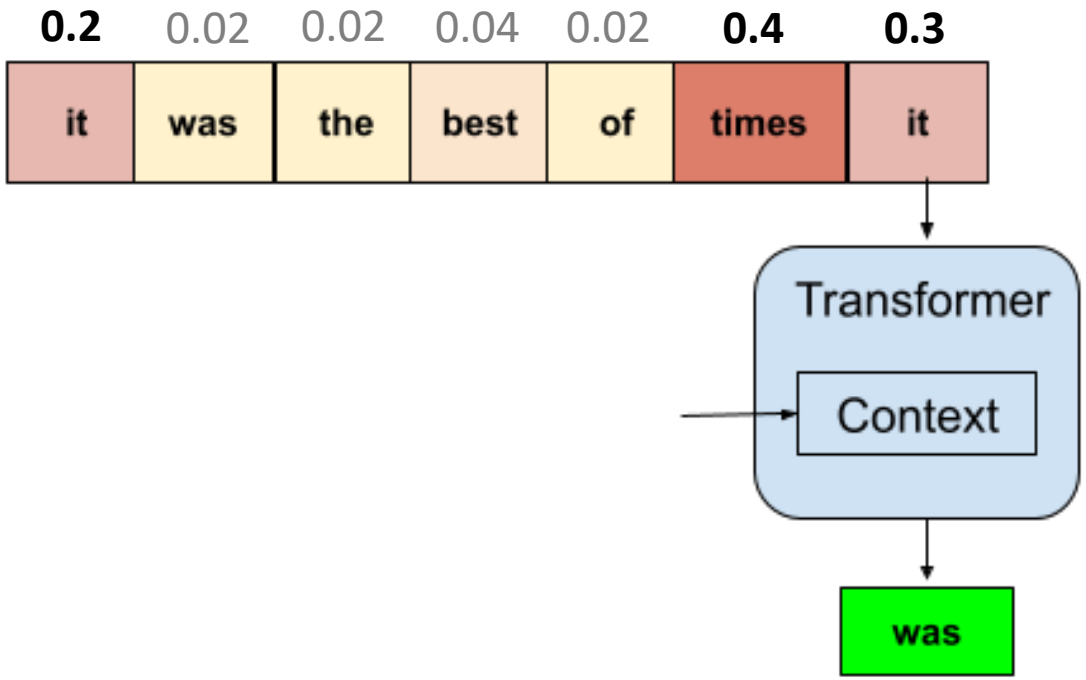
- *inference*: more “context” is better

Neural Language Modeling



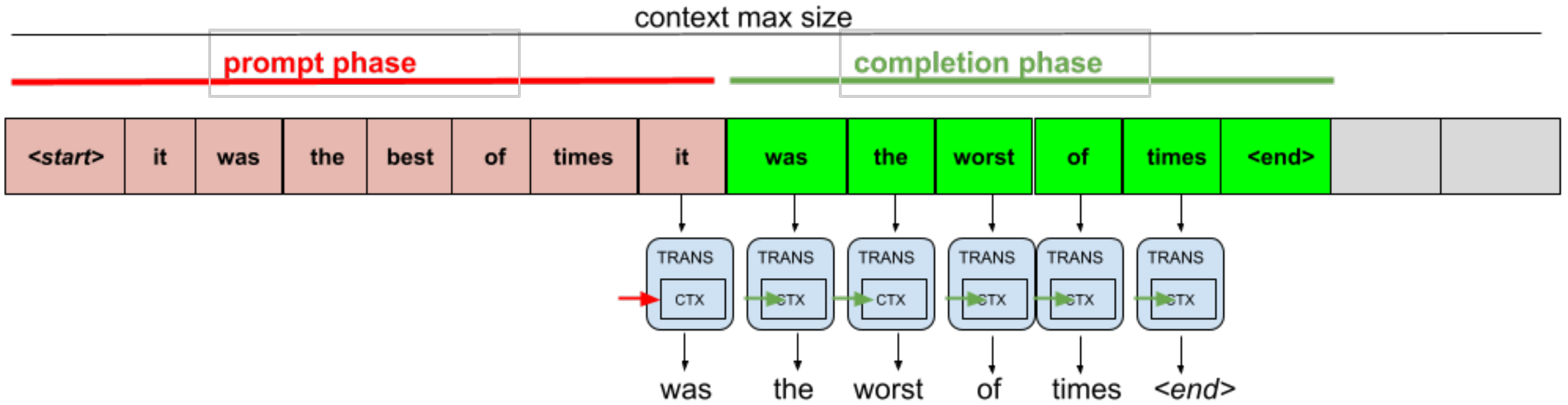
- *inference*: more “context” is better
- positional encoding

Neural Language Modeling

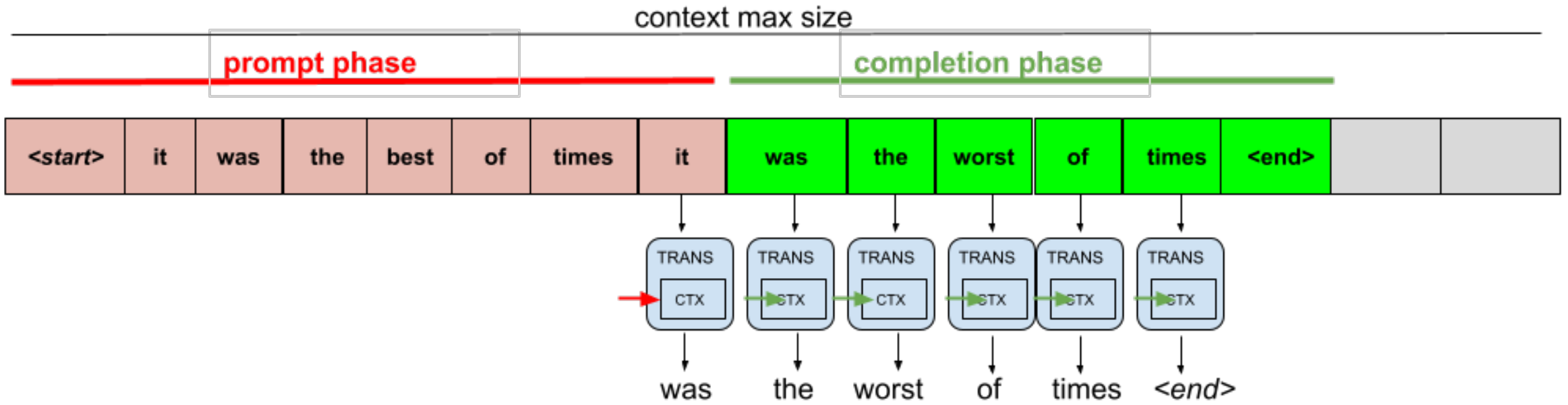


- *inference*: more “context” is better
- positional encoding
- *attention:* weighted focus

Neural Language Modeling



Neural Language Modeling



- *prompt phase*: tokens can be processed in parallel (compute bound)
- *completion phase:* tokens generated 1 at a time (IO bound)

I Asked ChatGPT…

Do Transformers Consume A Lot Of Power?

I Asked ChatGPT…

Do Transformers Consume A Lot Of Power?

Transformers, particularly large and deep ones like GPT-3 and its successors, can indeed consume a significant amount of power when running inference or training. The power consumption of a transformer model depends on several factors:

I Asked ChatGPT…

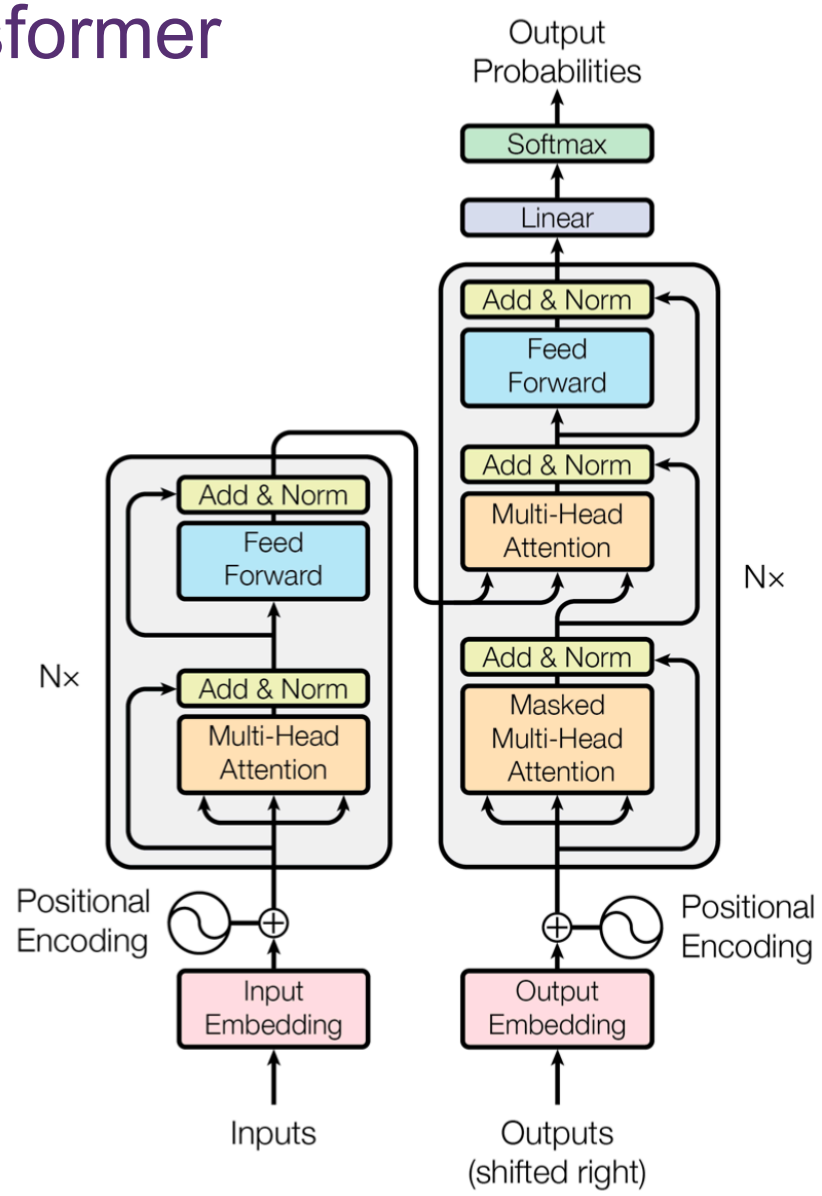
Do Transformers Consume A Lot Of Power?

Transformers, particularly large and deep ones like GPT-3 and its successors, can indeed consume a significant amount of power when running inference or training. The power consumption of a transformer model depends on several factors:

2. Hardware: The choice of hardware plays a crucial role in determining power consumption. Graphics processing units (GPUs) and application-specific integrated circuits (ASICs) designed for deep learning tasks can be more power-efficient than traditional central processing units (CPUs).

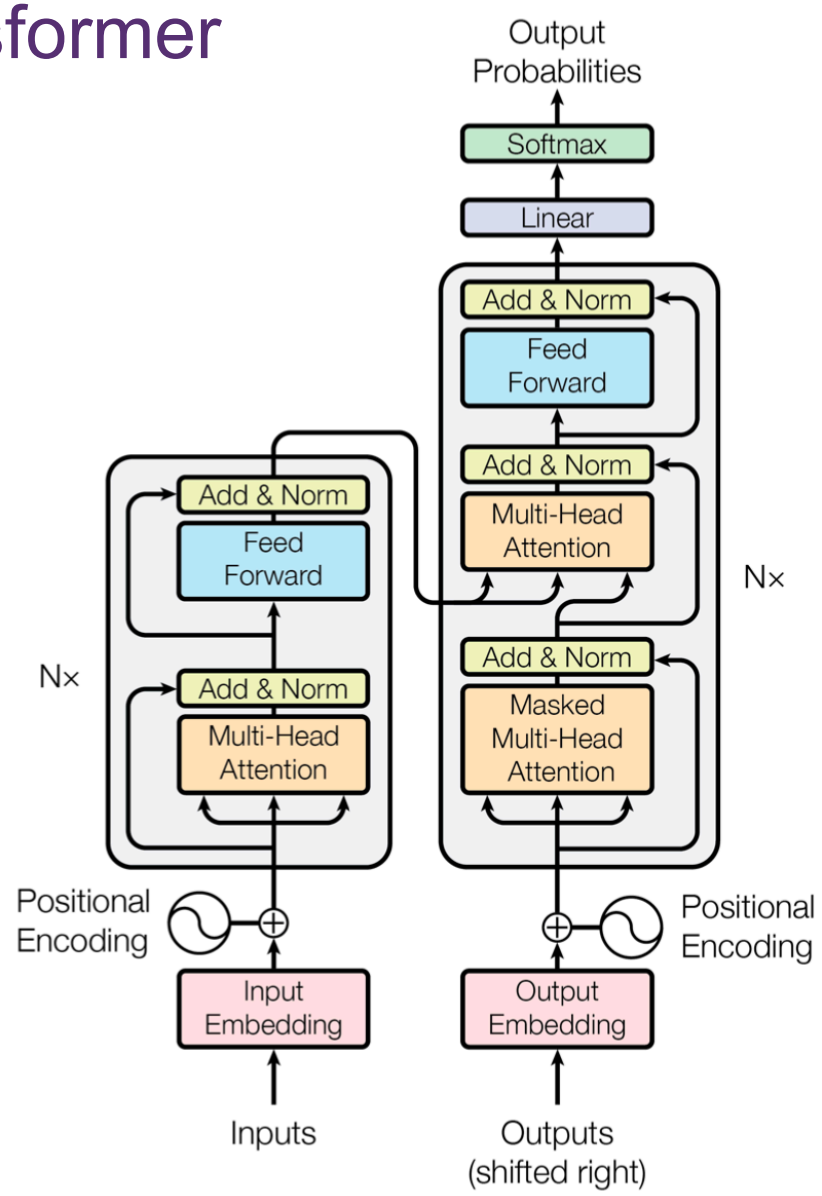
Transformer Essentials

Transformer



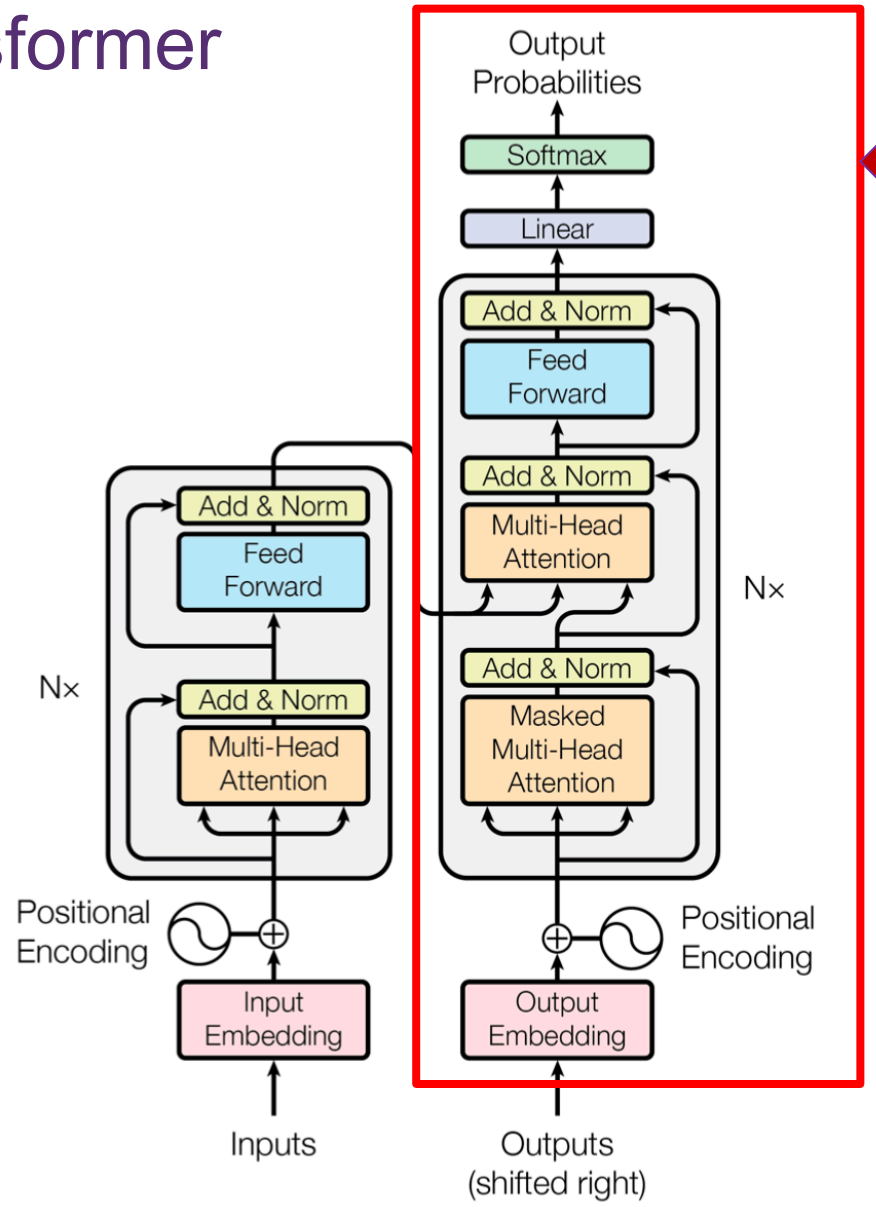
- 2017: “Attention Is All You Need”, Vaswani, et. al.

Transformer



- 2017: “Attention Is All You Need”, Vaswani, et. al.
- 2023: ChatGPT4, Llama 2, Palm 2, Claude 2, …
- OpenAI: 1 Billion in Revenue
- Nvidia: 100% YY Revenue

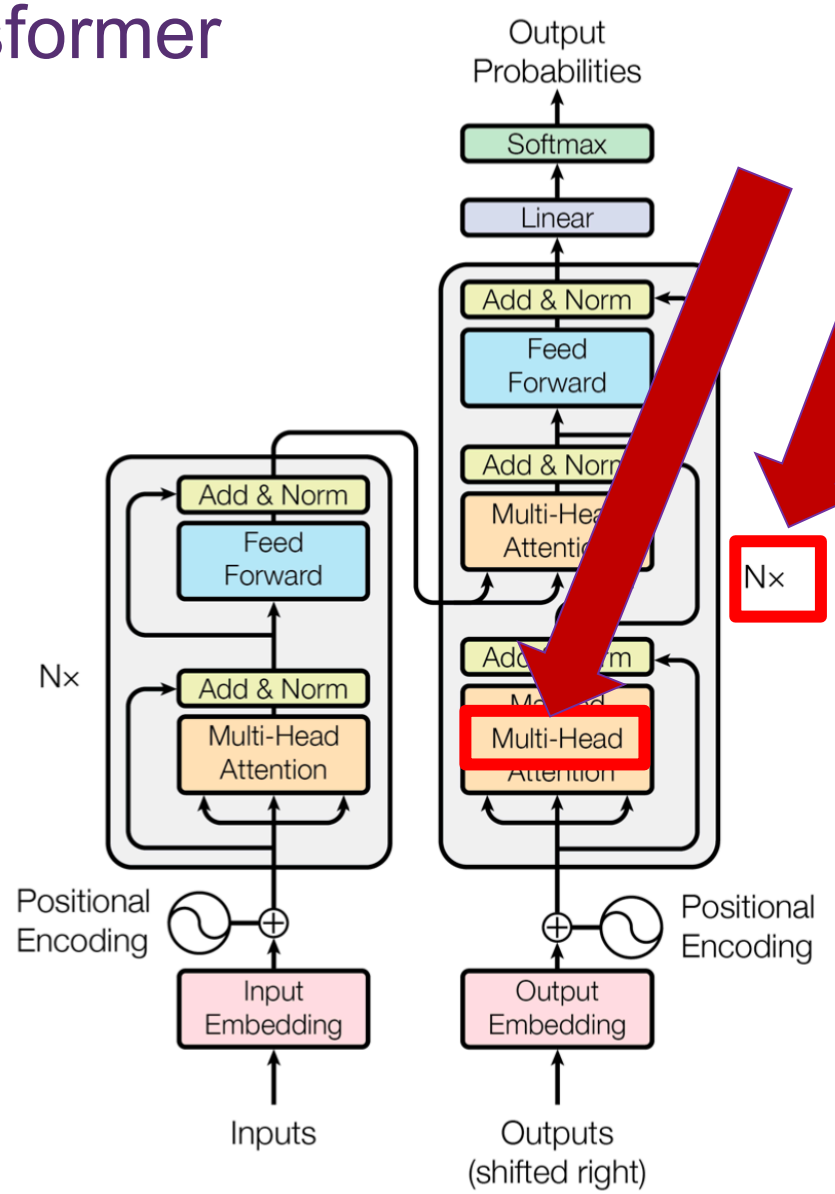
Transformer



Decoder Is All You Need!

- 2017: “Attention Is All You Need”, Vaswani, et. al.
- 2023: ChatGPT4, Llama 2, Palm 2, Claude 2, …
- OpenAI: 1 Billion in Revenue
- Nvidia: 100% YY Revenue

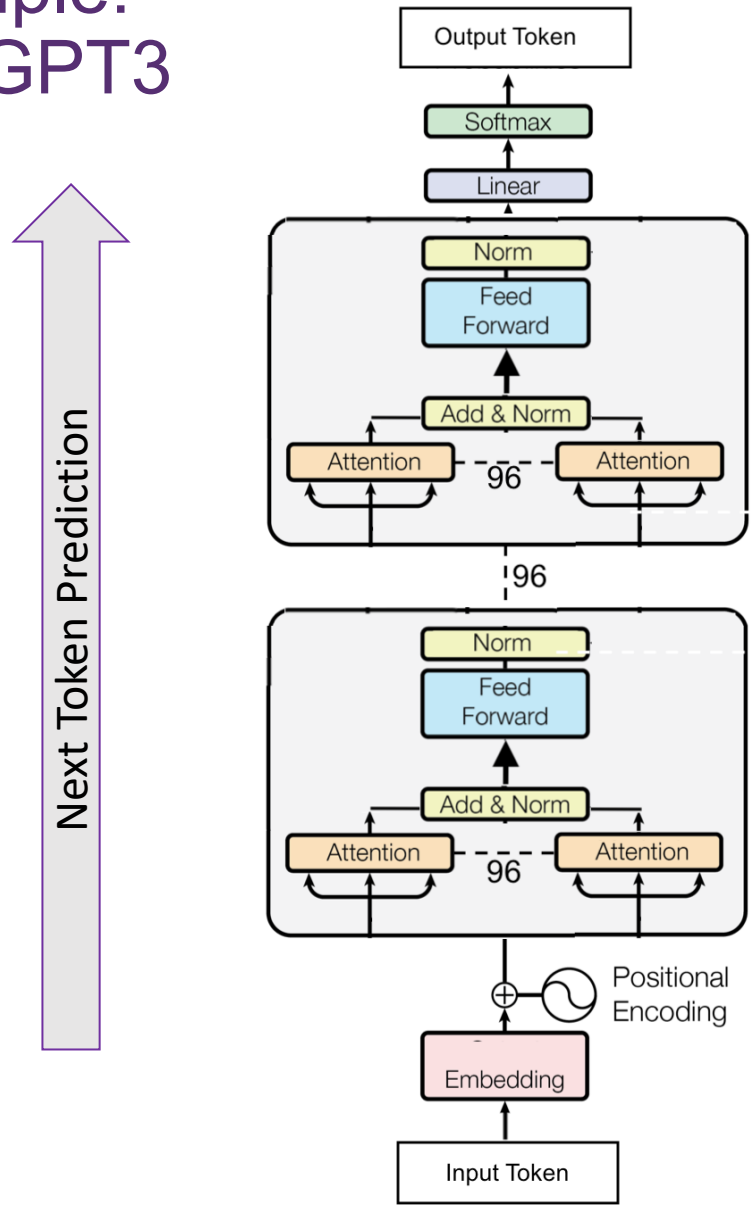
Transformer



Parameter Scaling!

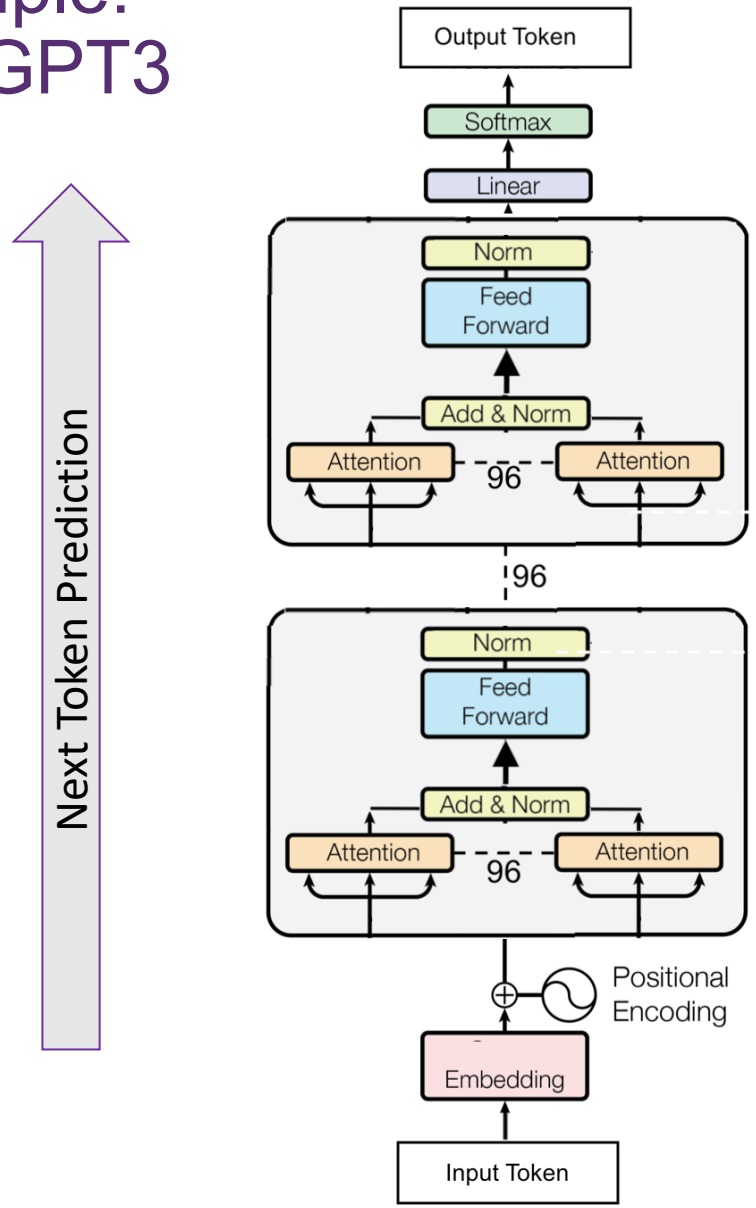
- 2017: “Attention Is All You Need”, Vaswani, et. al.
- 2023: ChatGPT4, Llama 2, Palm 2, Claude 2, …
- OpenAI: 1 Billion in Revenue
- Nvidia: 100% YY Revenue

Example: ChatGPT3



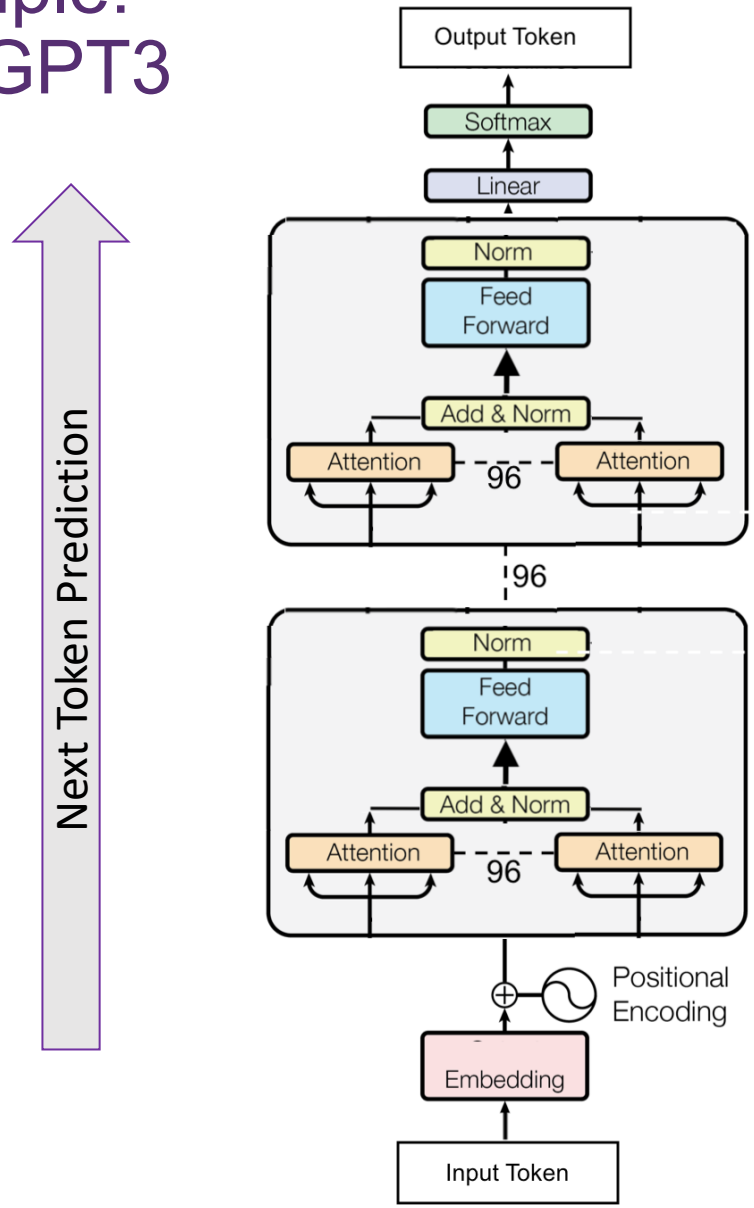
- 96 layers
- 96 “attention heads”

Example: ChatGPT3



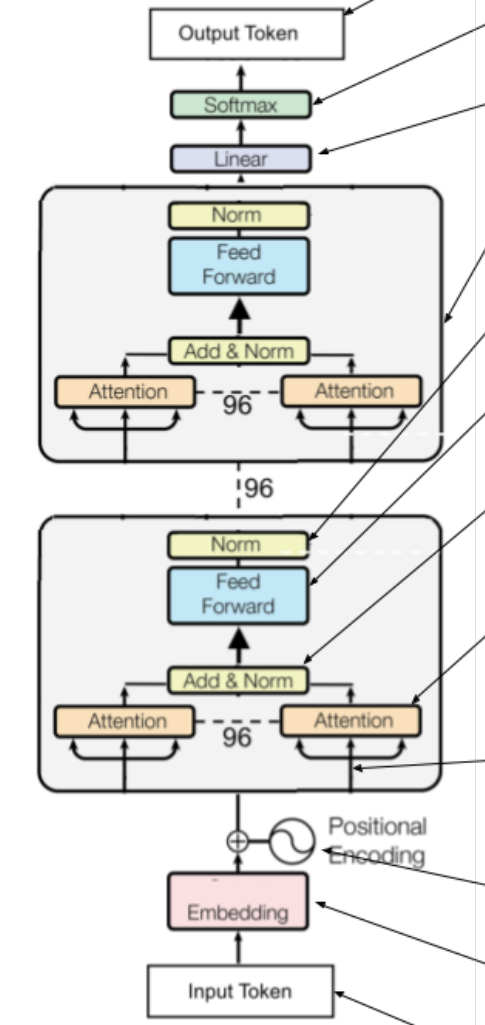
- 96 layers
- 96 “attention heads”
- 175 billion parameters (“weights”)

Example: ChatGPT3



- 96 layers
- 96 “attention heads”
- 175 billion parameters (“weights”)
- Training from scratch requires weeks on 10s-100s of GPUs

Example: ChatGPT3



Output Token: Produce the actual token (word, punctuation, etc.)

Softmax: A operation which produces 1-hot vector (vocab len).

Linear: A matmul from 12288-D to 50K-D (vocab len)

Repeat Decoder Block: Repeat the decoder layer operations 95 more times (96 total)

Norm: Layer normalization is applied.

Feed Forward: The 12288-D vector is fed into a 2-layer FC DNN

Add&Norm: Each head produces a 128-D vector which gets concatenated to 12288-D and then layer normalized.

Attention: Each head applies "attention" on ALL TOKENS seen so far sequence leveraging the the context state (called KV cache) to do so.

Multi-Head Attention Projection: Each "attention head" (96 total) takes the 12288-D vector and projects to 3 128-D (called query, key, value) via matrix multiplication.

Pos Enc: Add a constant vector based on position in the sequence

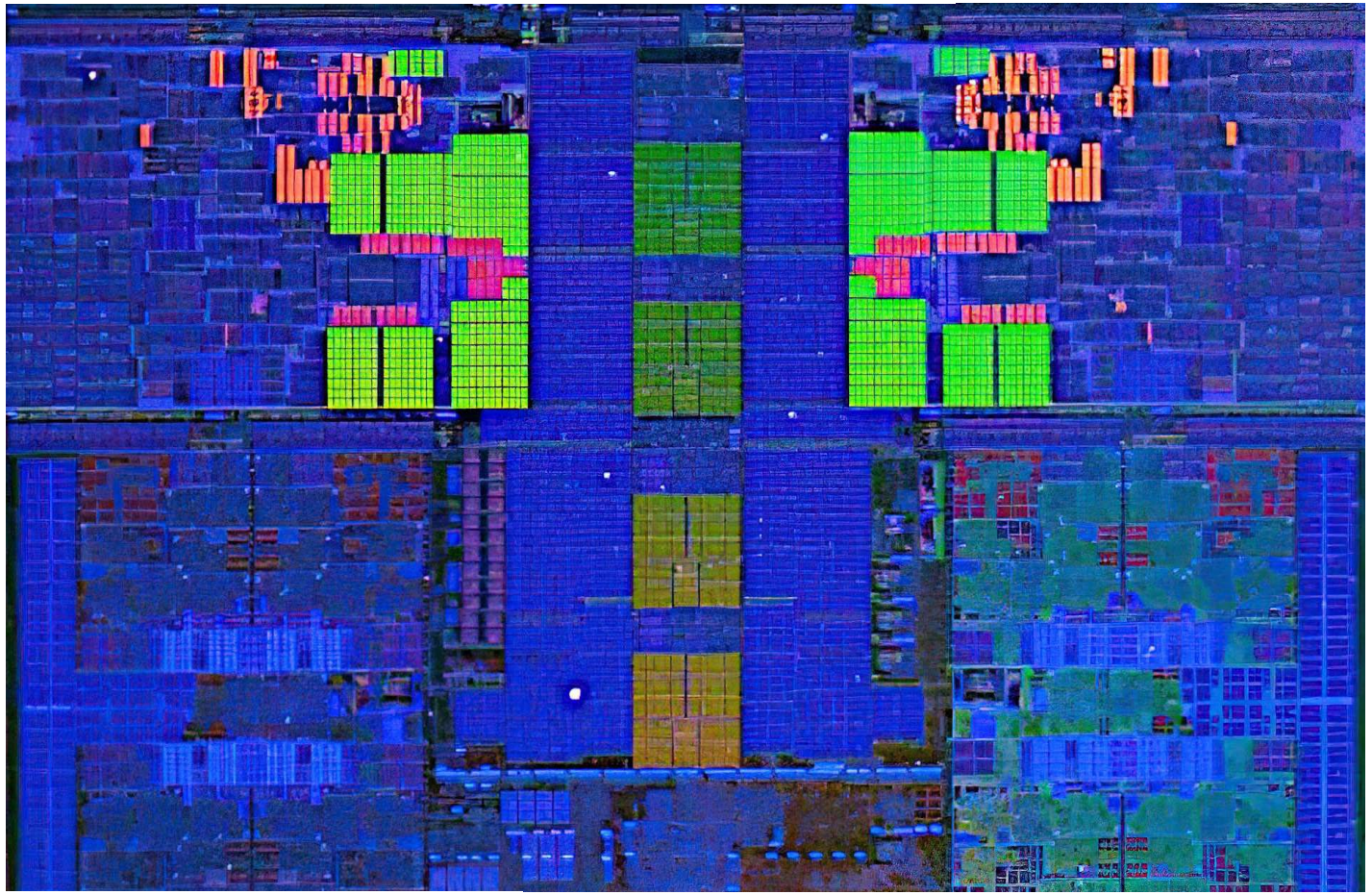
Embedding: Converts 1-hot token to a vector dim=12288

Input Token: 1 hot encoding of the token to 50K-D vector (vocab len)

Compute-In-Memory For Transformer

Typical Von-Neumann Architecture

The dominate compute paradigm for 60 years!

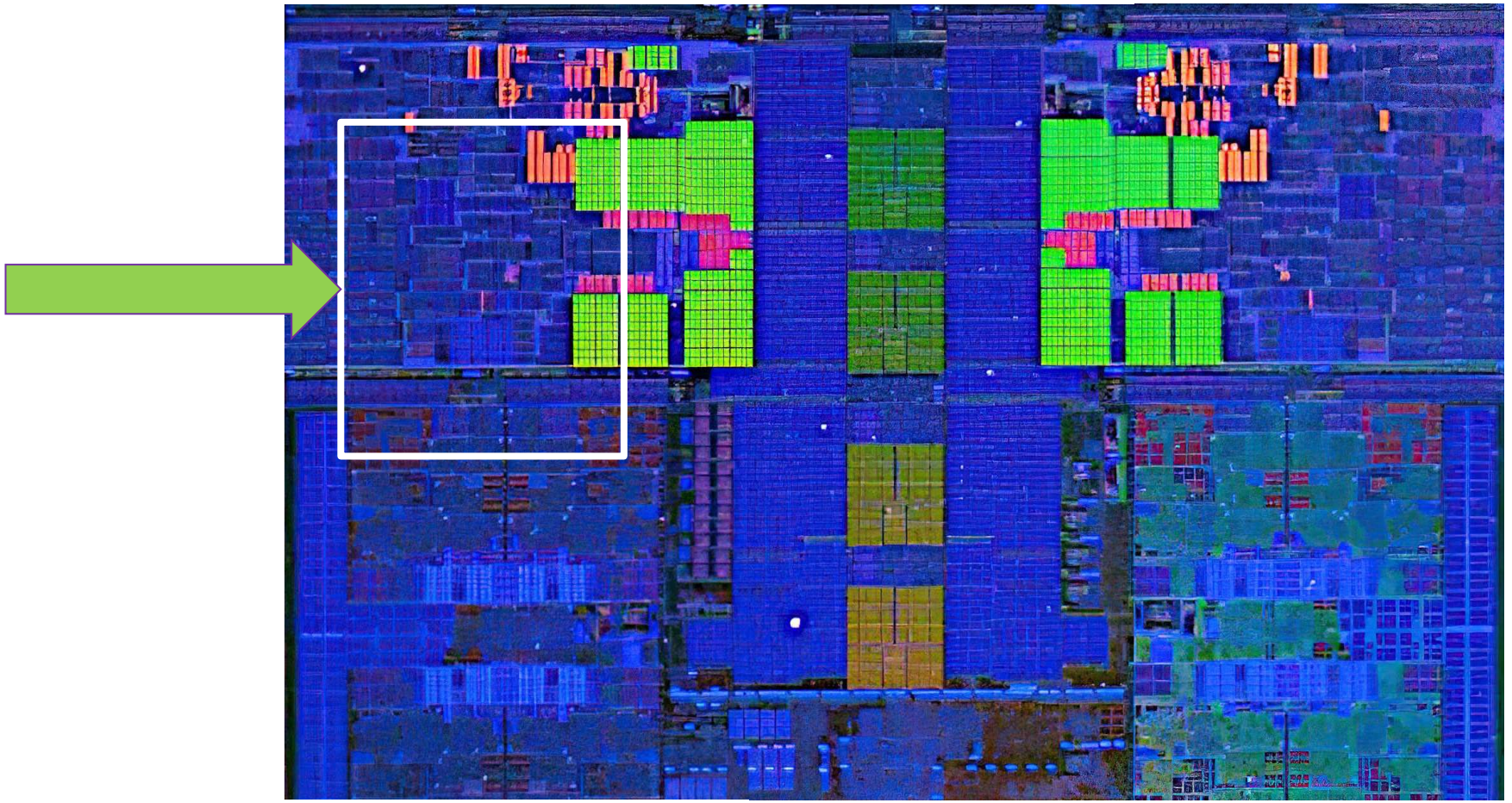


Intel Meteor Lake Die

Where is memory and where is compute?

Typical Von-Neumann Architecture

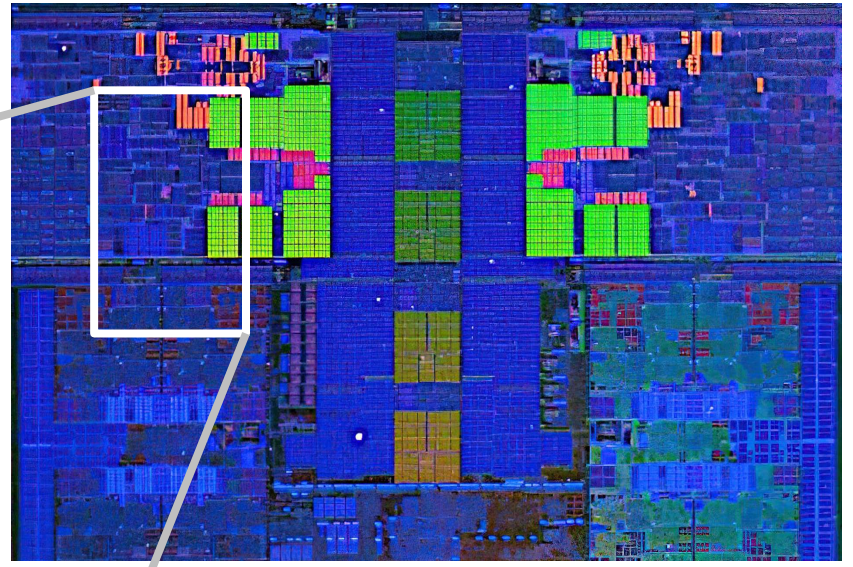
The dominate compute paradigm for 60 years!



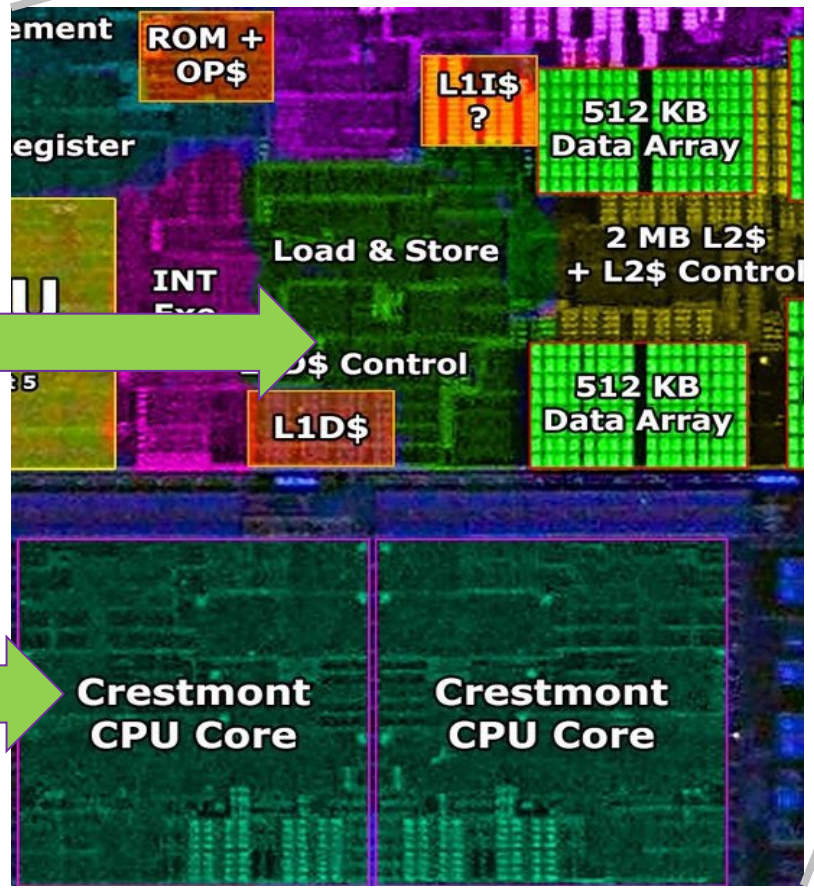
Intel Meteor Lake Die

Typical Von-Neumann Bottleneck

The dominate compute paradigm for 60 years!



Intel Meteor Lake Die



L1 SRAM

Compute Core

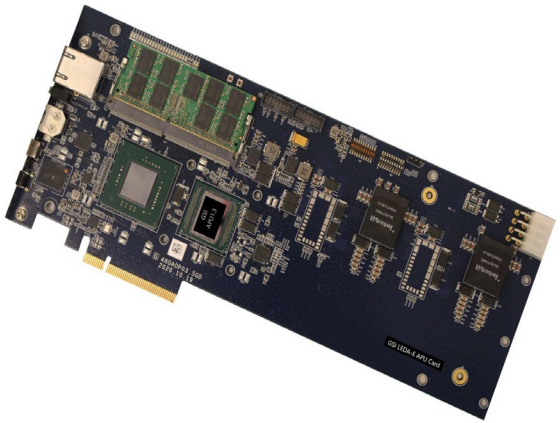
Is there a better way?

In-Memory-Computing Hardware Landscape

RAIN

Digital IMC

Analog IMC

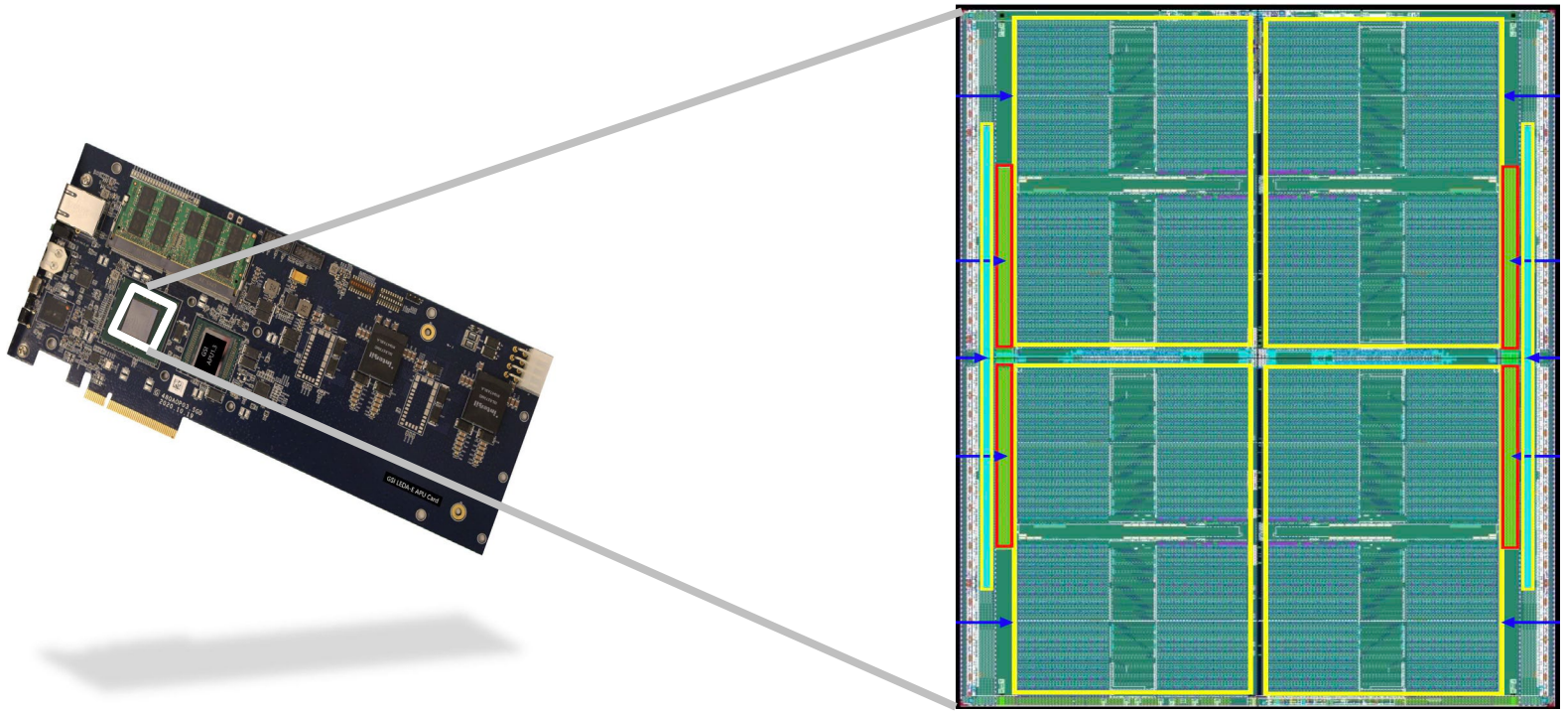


GSI TECHNOLOGY

Associative IMC

SDC 23

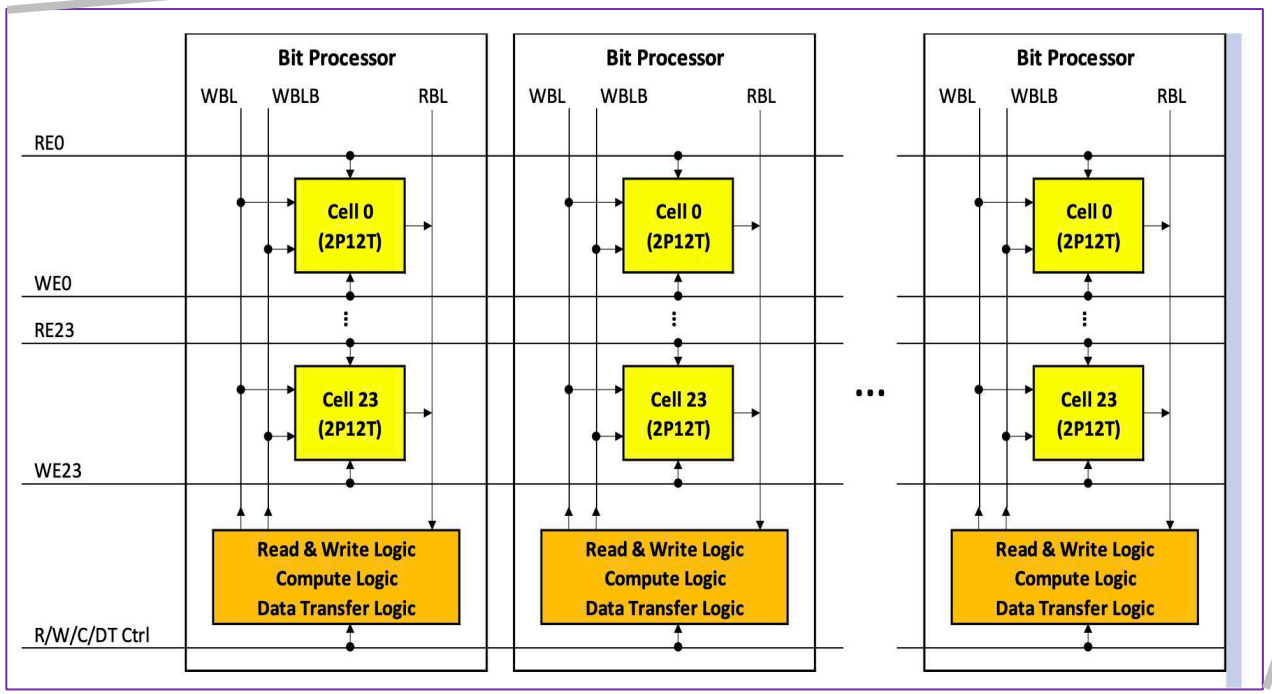
GSI Technology's Associative Processor

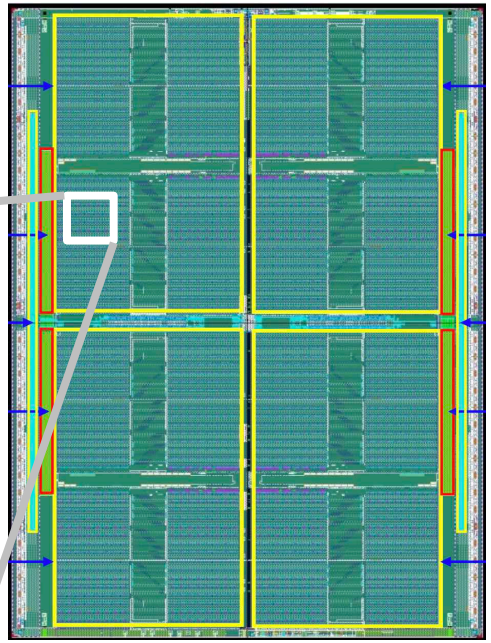


GSI APU (G1)

Add Processors Into SRAM

Compute-in-Memory paradigm…



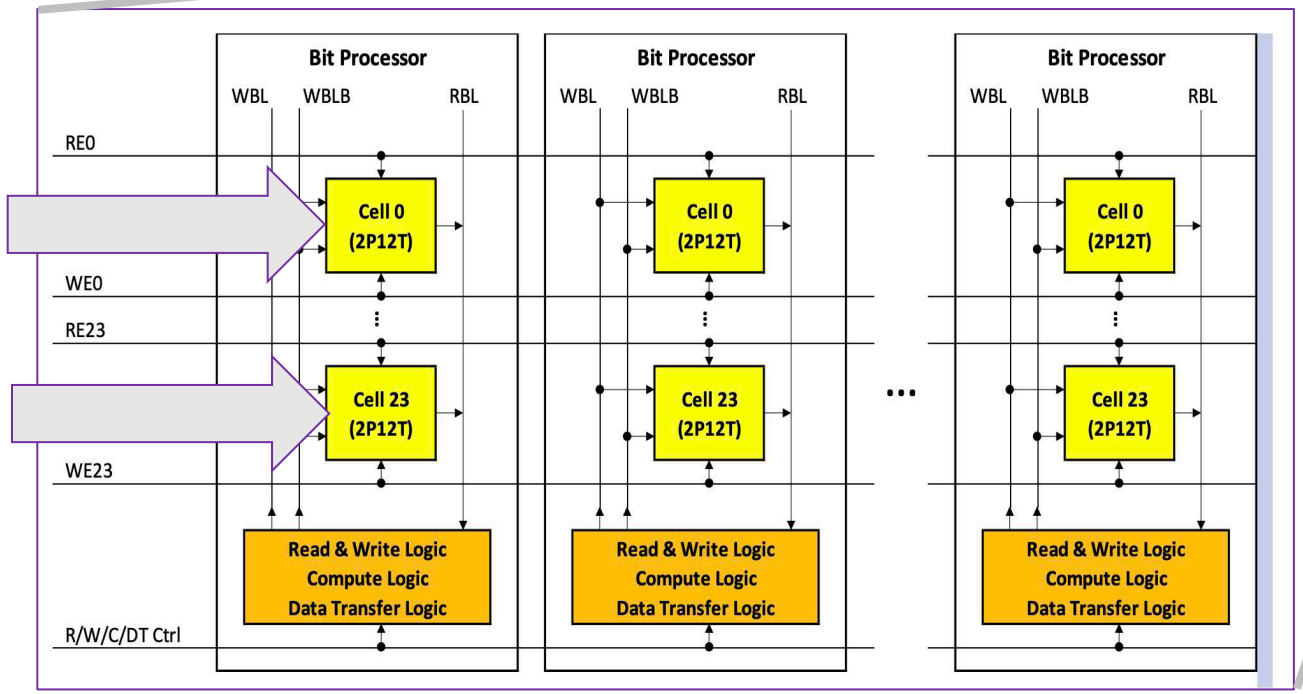


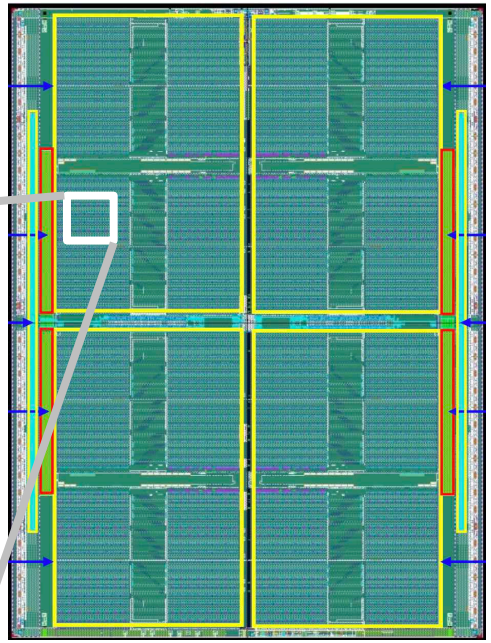
GSI APU (G1)

Add Processors Into SRAM

A “typical” SRAM grid…

SRAM Cells



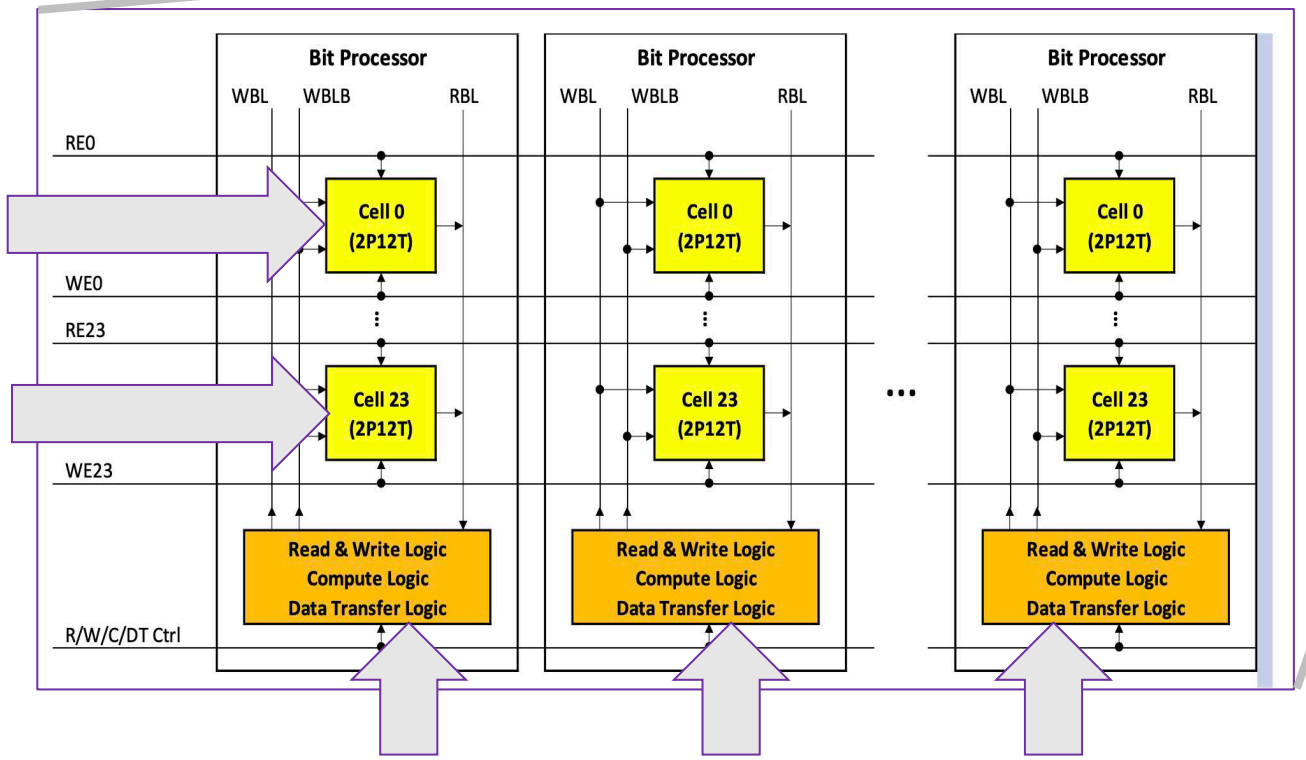


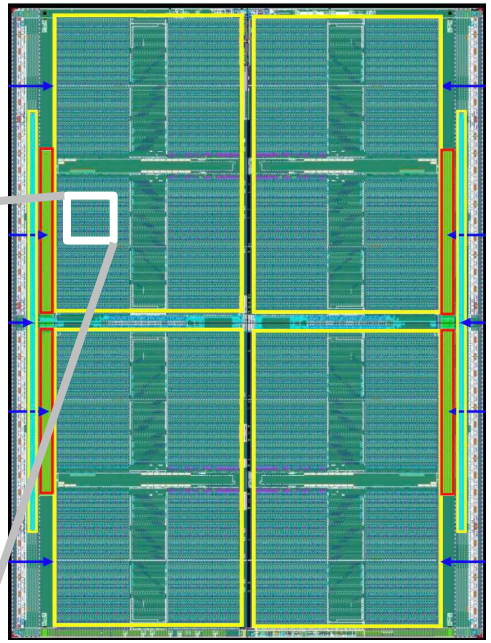
GSI APU (G1)

Add Processors Into SRAM

A “typical” SRAM grid with interleaved processors.

SRAM Cells





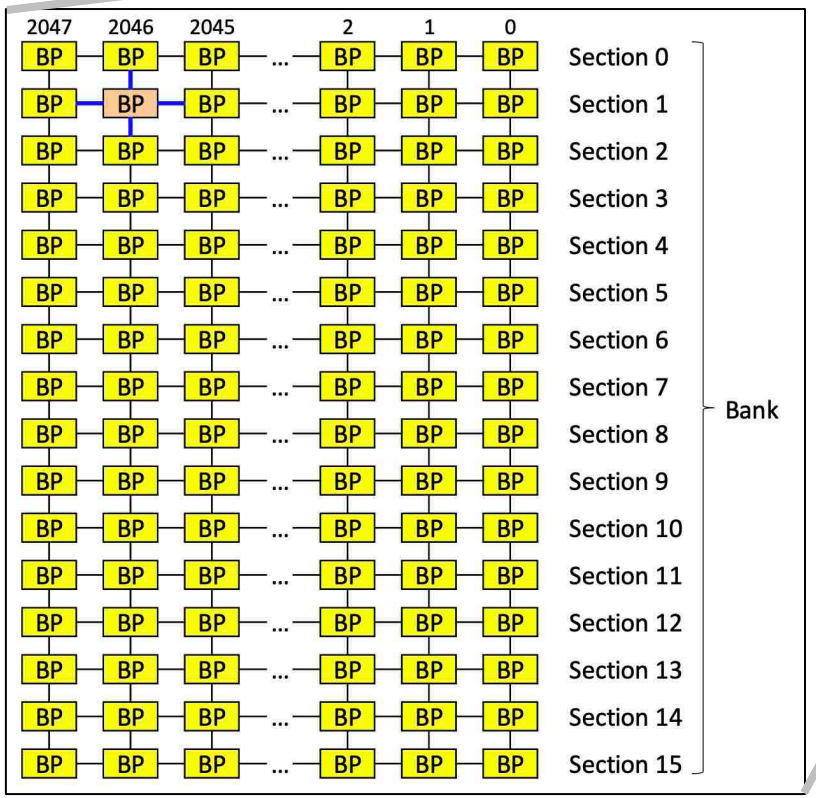
GSI APU (G1)

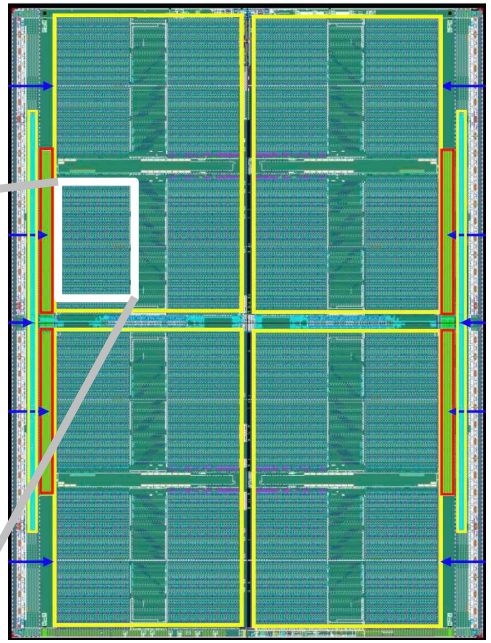
20 microns (avg) between BP and SRAM

Bit Processors (BPs) are fully parallel and programmable

Bit Processors

Associative Processing

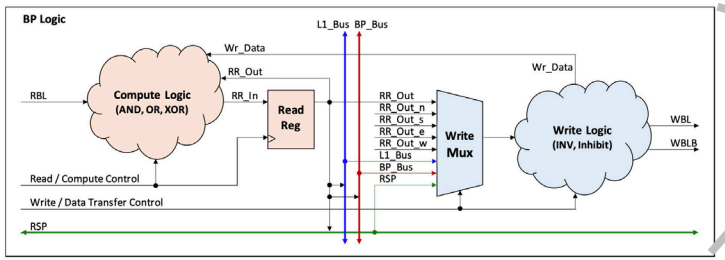


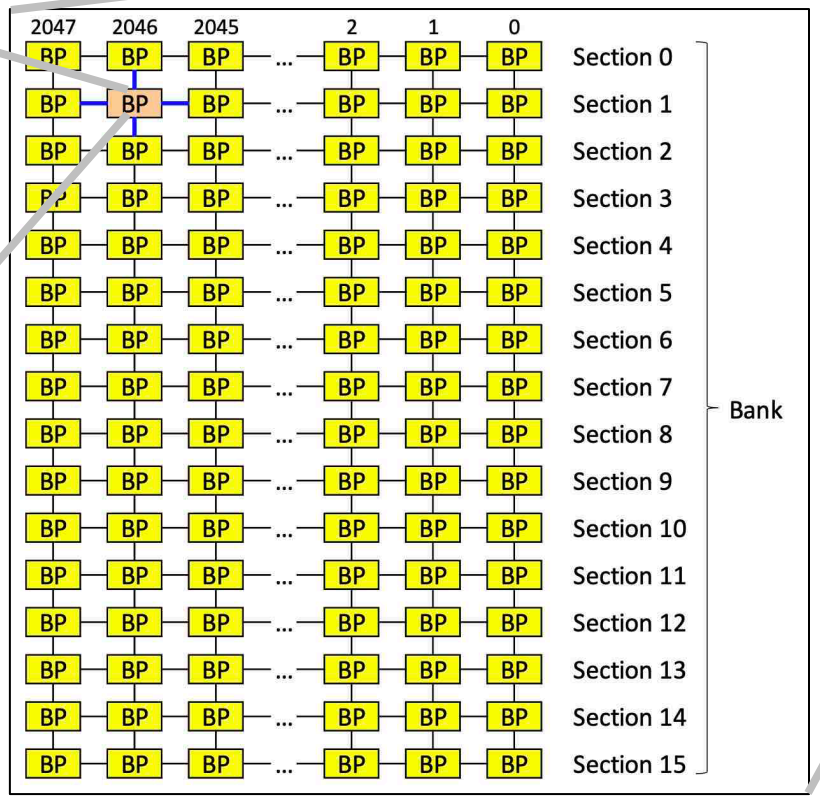


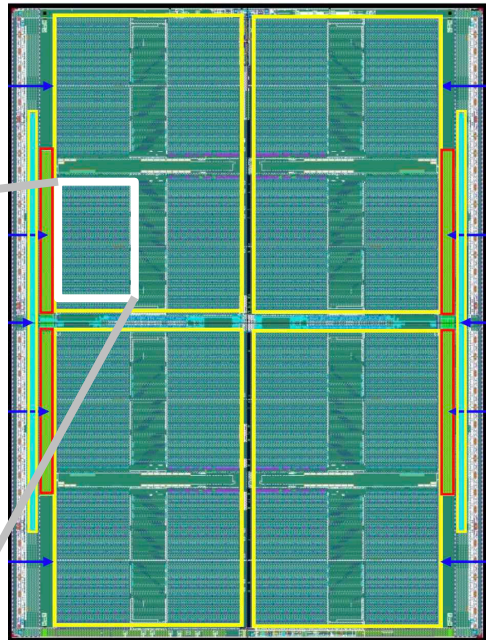
GSI APU (G1)

Associative Processing

Each BP is Simple…



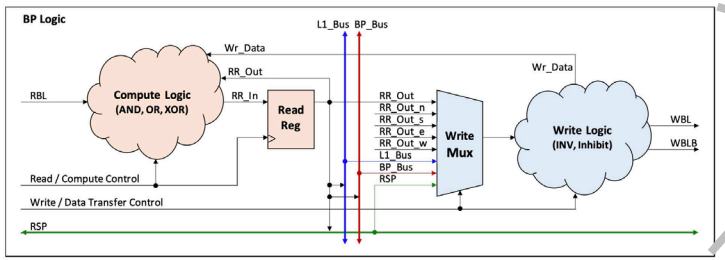


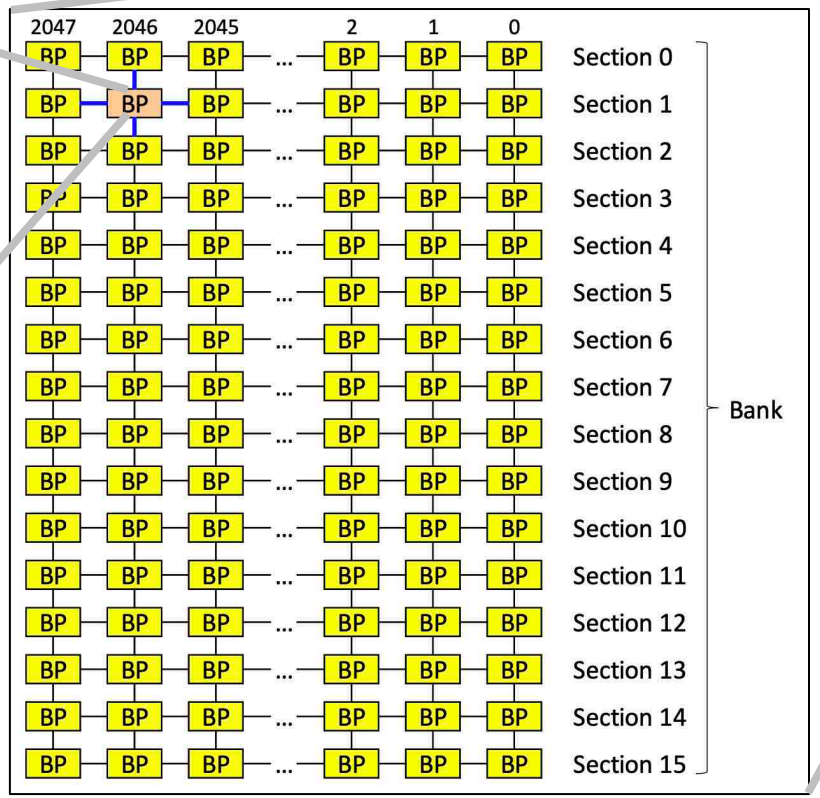


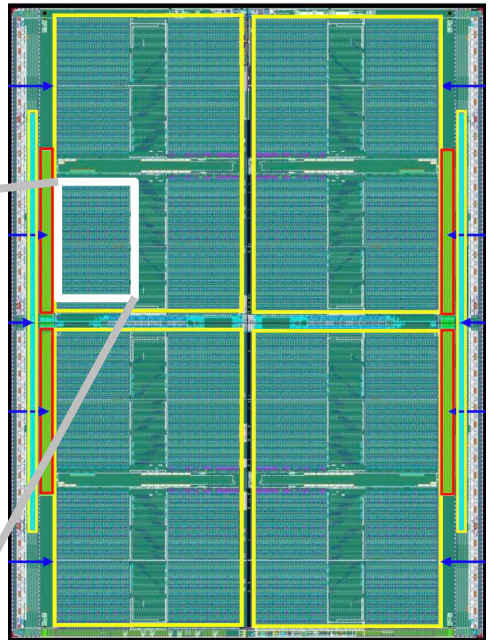
GSI APU (G1)

Associative Processing

***MxN BPs forms a powerful* compute grid (2M<->48Mb)**

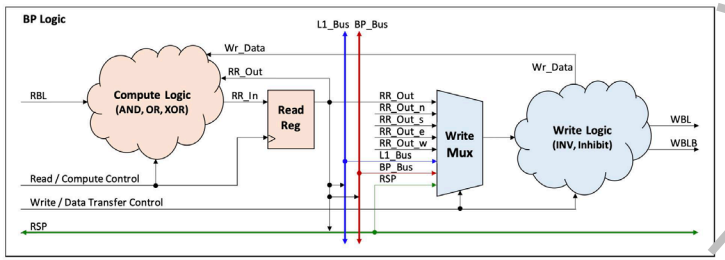


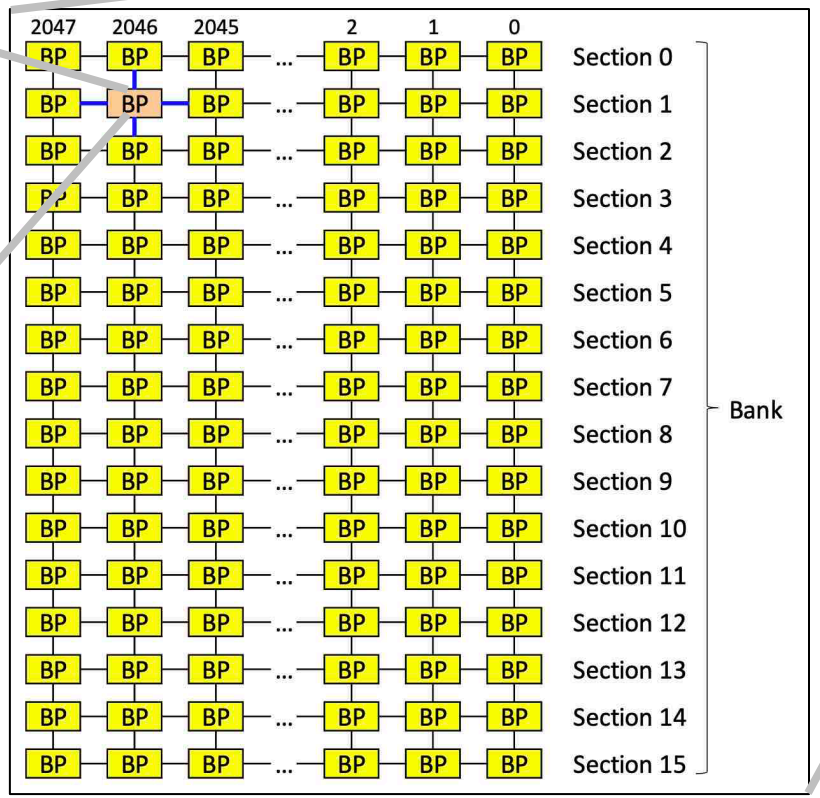


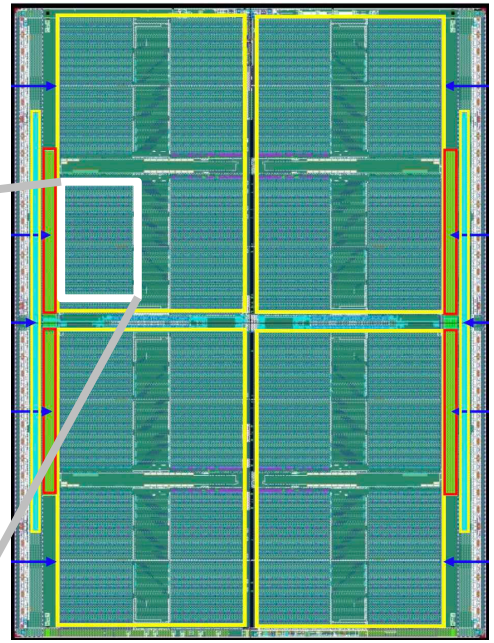


GSI APU (G1)

Associative Processing

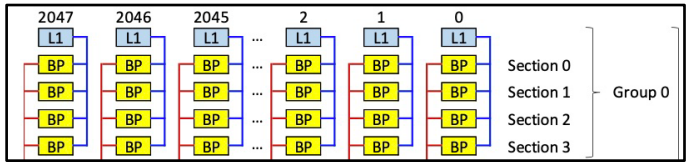






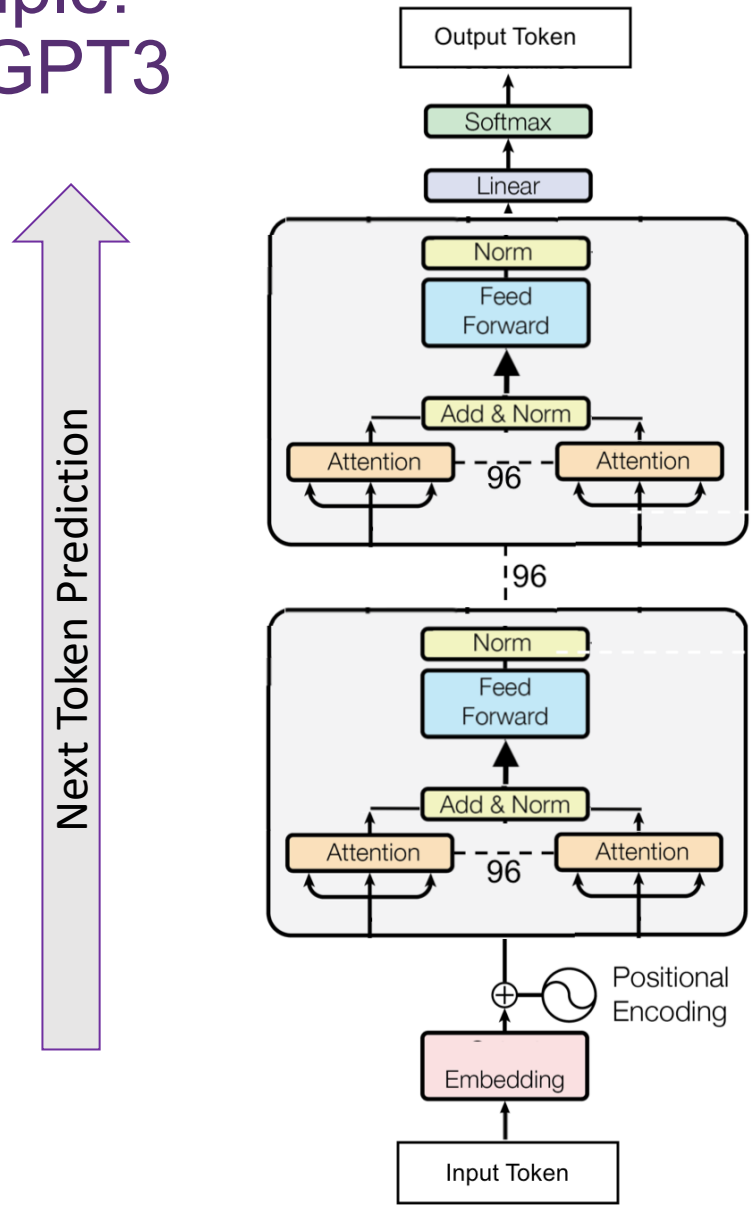
GSI APU (G1)

L1 is interleaved too (96Mb)

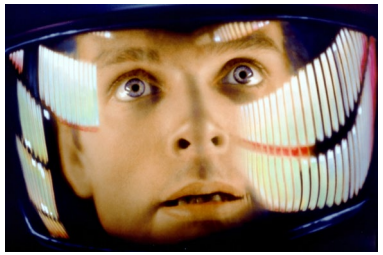


100 microns (avg) between BP and L1

Example: ChatGPT3

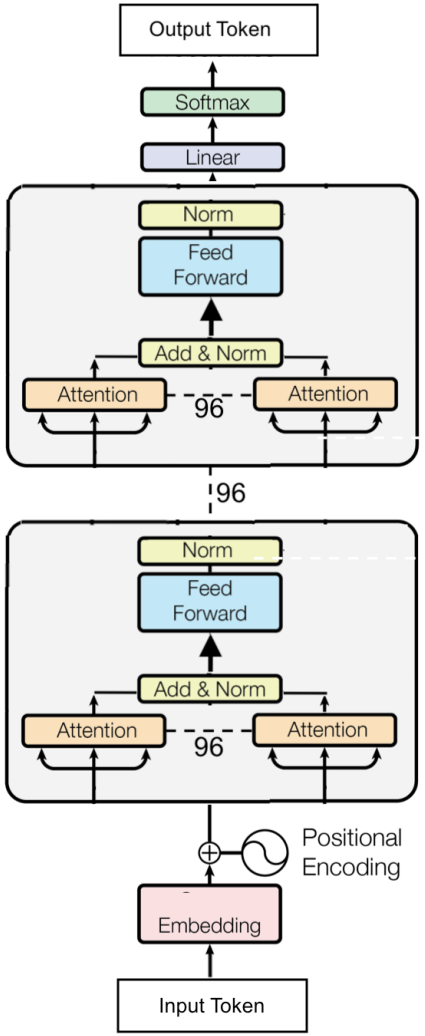


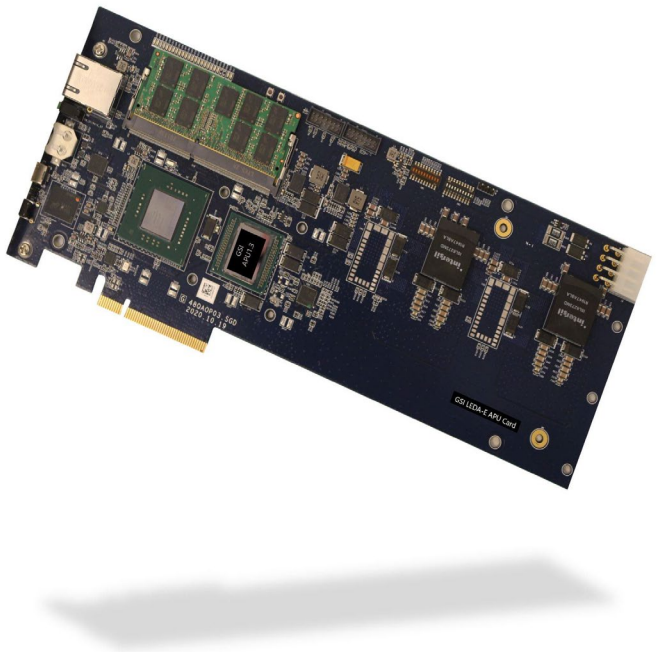
- 96 layers
- 96 “attention heads”
- 175 billion parameters (“weights”)
- **Most operations are MAC for matrix multiplication**



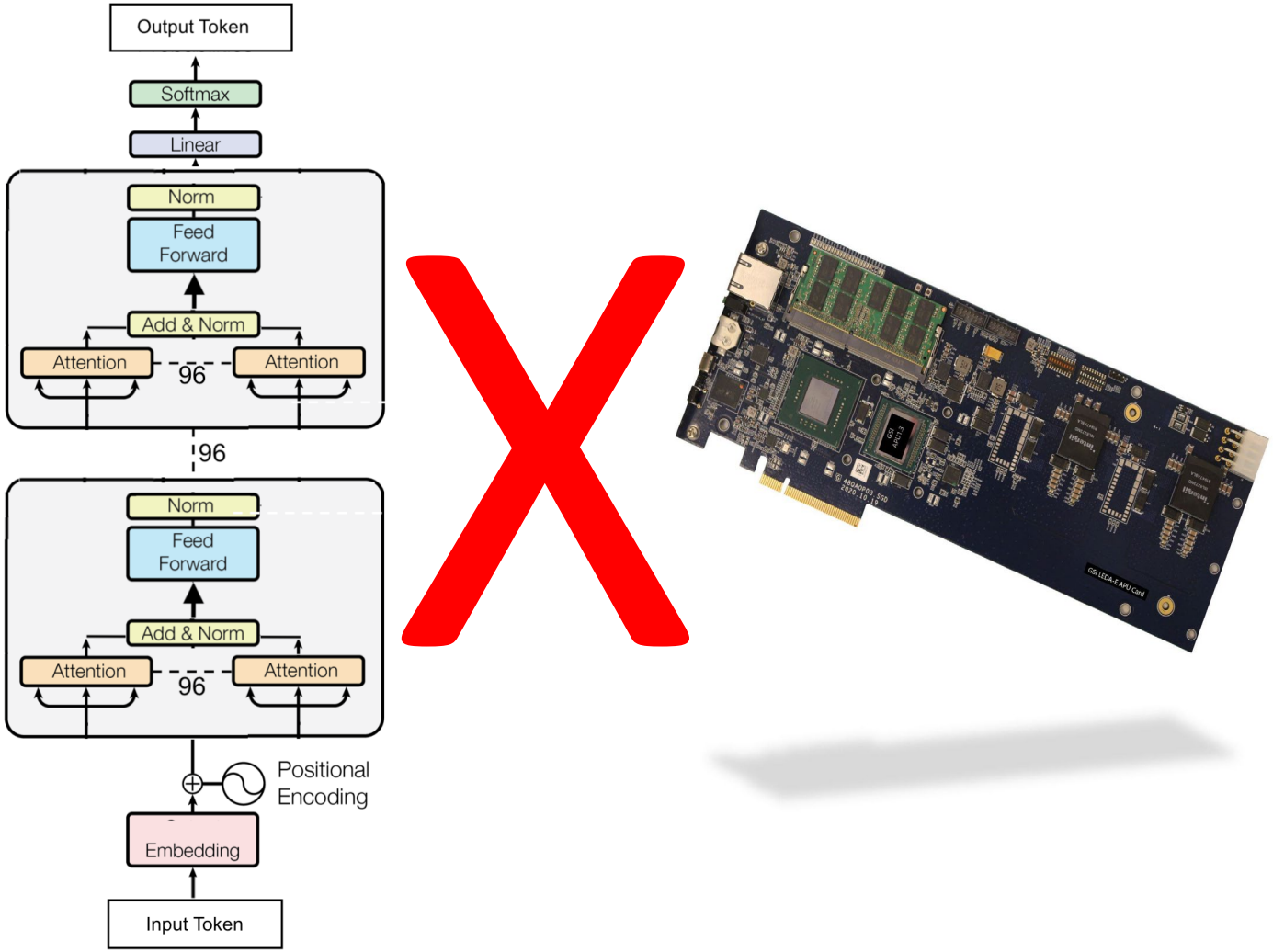
“…it’s full of ~~stars~~ MatMul!”

Low Power LLM?





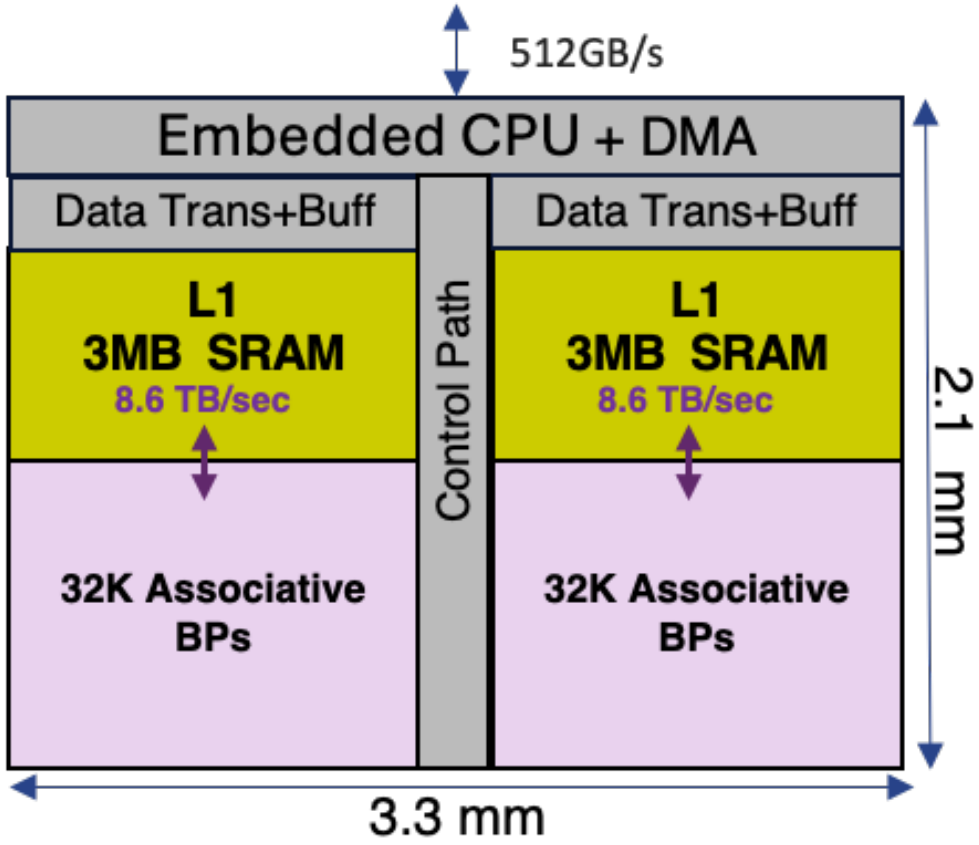
Low Power LLM?



Modular IP For Reticle and Power Budgets

Example: MatMul "Tiling" with 6MB

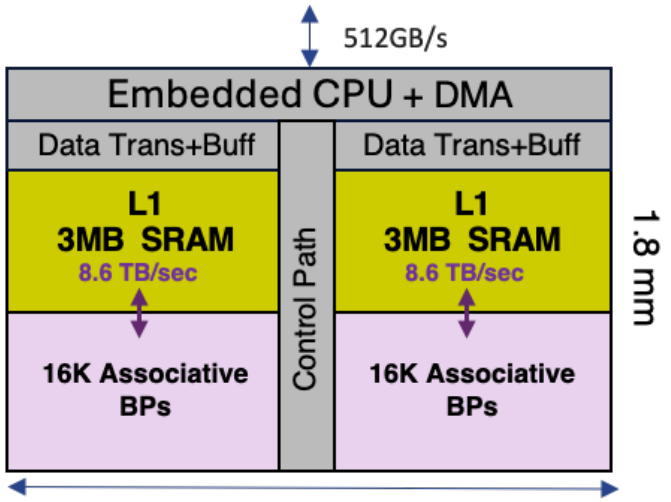
6.1 TOPS (INT8)
4.8 TOPS (FP8)
5W TDP



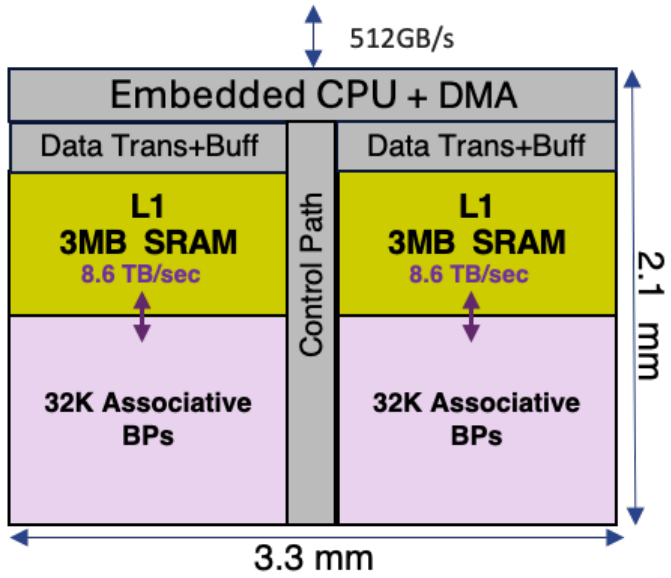
Memory "bank" architecture accommodates different size and power profiles…

Modular IP For Reticle and Power Budgets

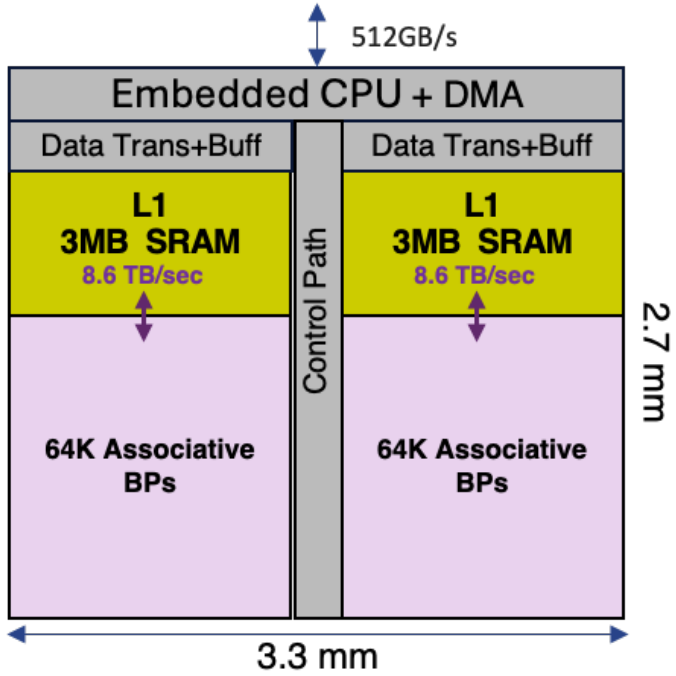
Example: MatMul "Tiling" with 6MB



3.1 TOPS (INT8)
2.4 TOPS (FP8)
2.7W TDP

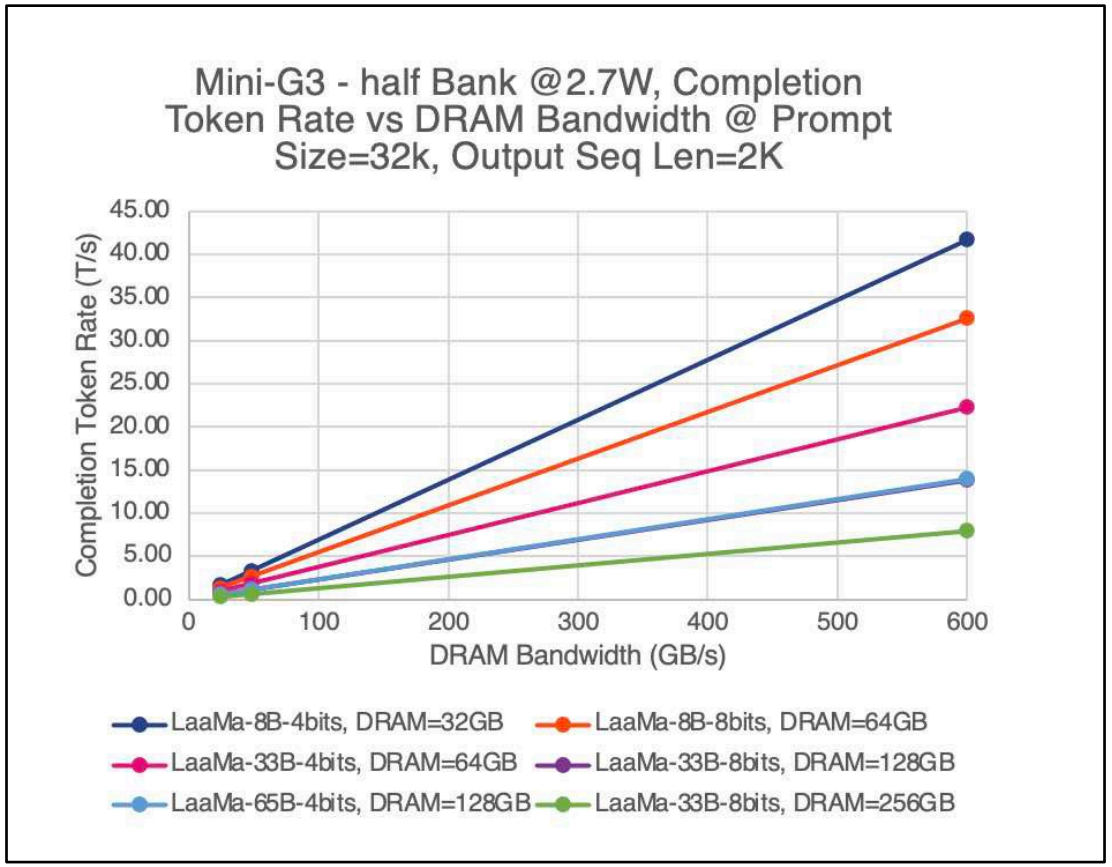


6.1 TOPS (INT8)
4.8 TOPS (FP8)
5W TDP



12.2 TOPS (INT8)
9.5 TOPS (FP8)
10W TDP

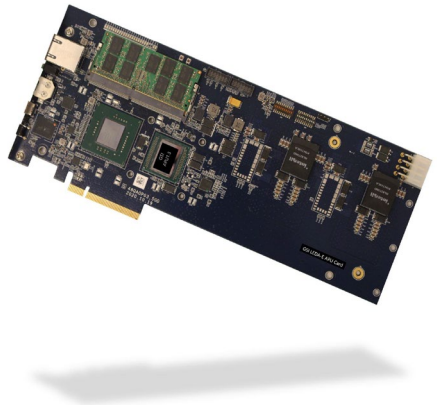
Llama2 Completion Phase Token Rates



Try It Out!

Product	*Avail*
G1 with / 2M BPs in PCIe	Now
G2 with / X10 L1 Interleaved Cache	Q4
Microcode Compiler For C/Python (OSS)	Now
Modular IP Licensing	Q4

GSI TECHNOLOGY



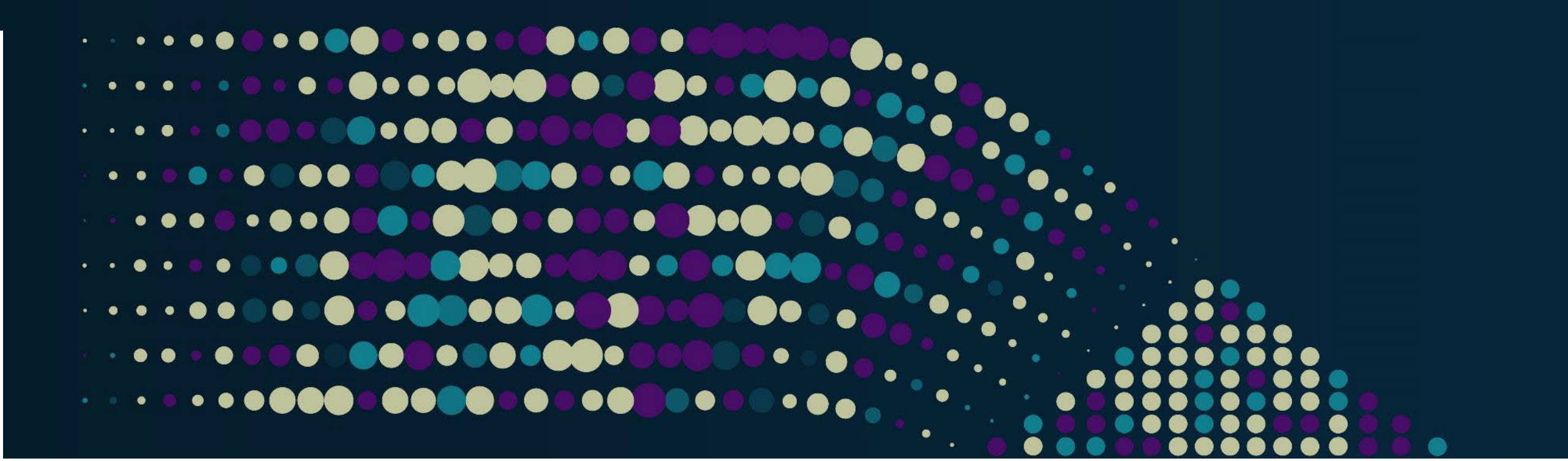
associativecomputing@gsitechnology.com

SDC 23

The End

Section Title

Section Subtitle



Section Title

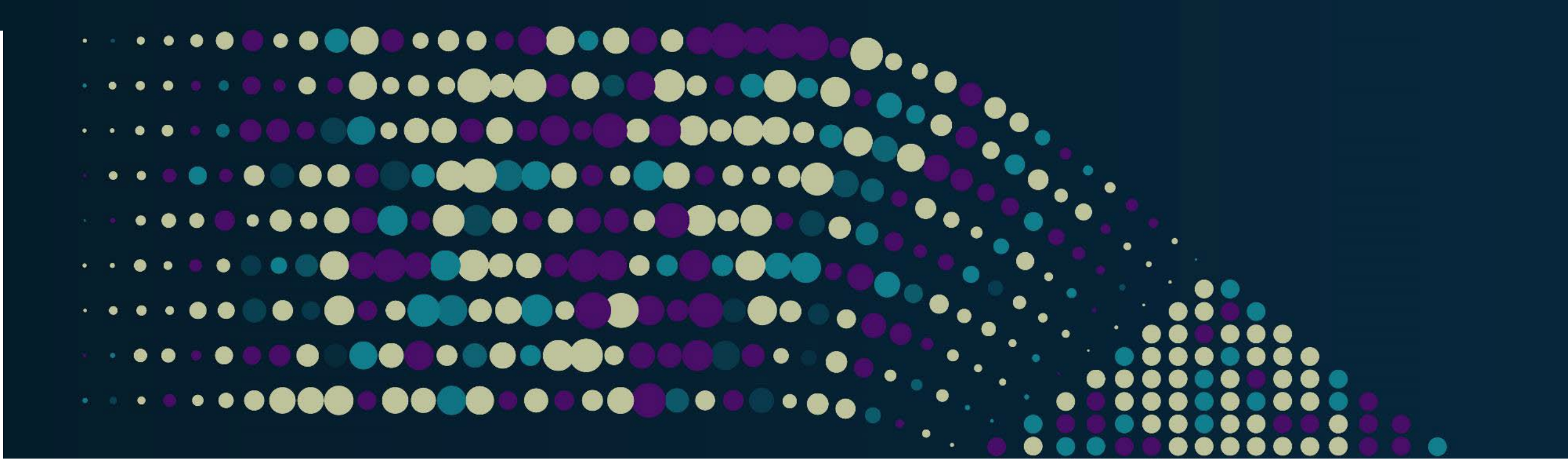
Section Subtitle

Light Slide Title

- Bullets 1
  - Bullets 2
    - Bullets 3
      - Bullets 4
        - Bullets 5

Dark Slide Title

- Bullets 1
  - Bullets 2
    - Bullets 3
      - Bullets 4
        - Bullets 5



Please take a moment to rate this session.

Your feedback is important to us.

