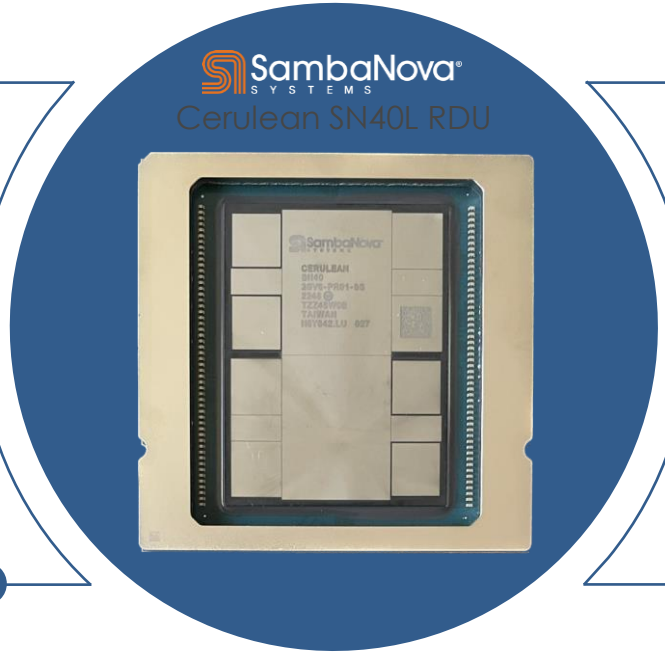
SambaNova SN40L RDU:
Breaking the Barrier of Trillion+ Parameter Scale Gen AI Computing

Raghu Prabhakar
Architect, SambaNova Systems

HotChips 2024

SN40L: SambaNova's New Language-Optimized RDU

"Cerulean" Architecture-based Reconfigurable Dataflow Unit



5nm TSMC

102B Transistors

1,040 RDU Cores

638 TFLOPS (bf16)

3-tier Dataflow Memory

520 MB
On-Chip Memory

64 GB
High Bandwidth Memory

1.5 TB
High Capacity Memory

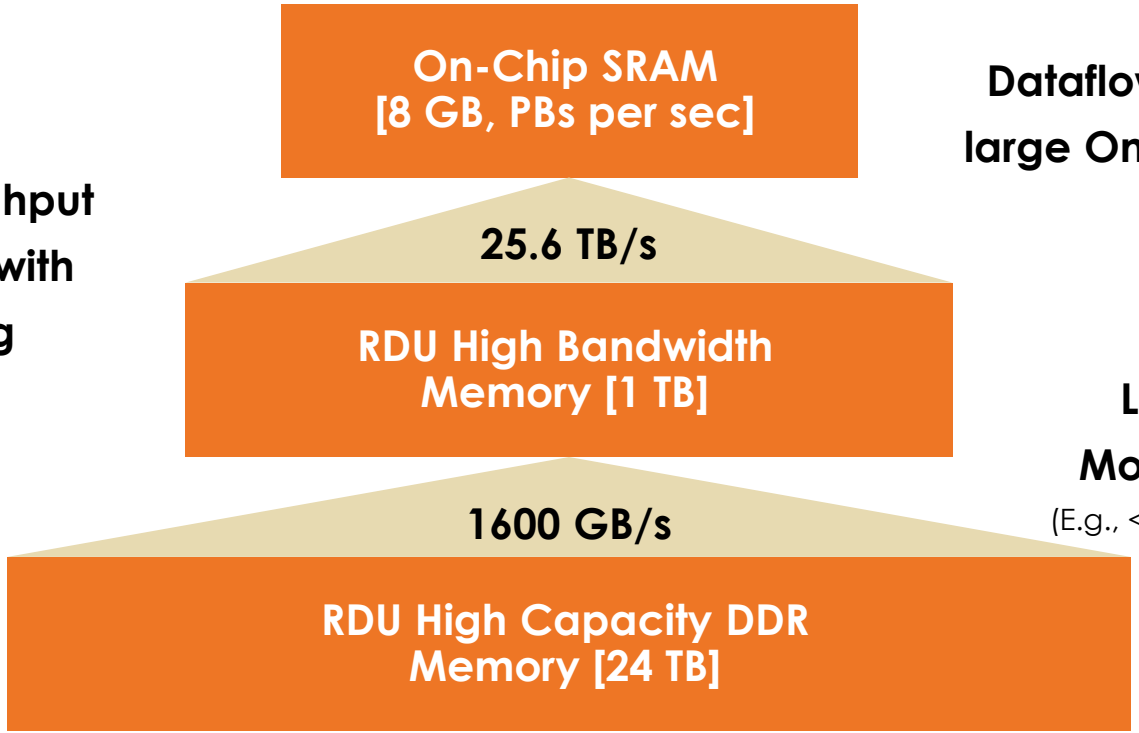
Generative AI Training and Inference

SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

SN40L: SambaNova's New Language-Optimized RDU

3-tier Memory System with SRAM, HBM, and DDR



High throughput inference with caching

Dataflow enabled by large On-Chip Memory

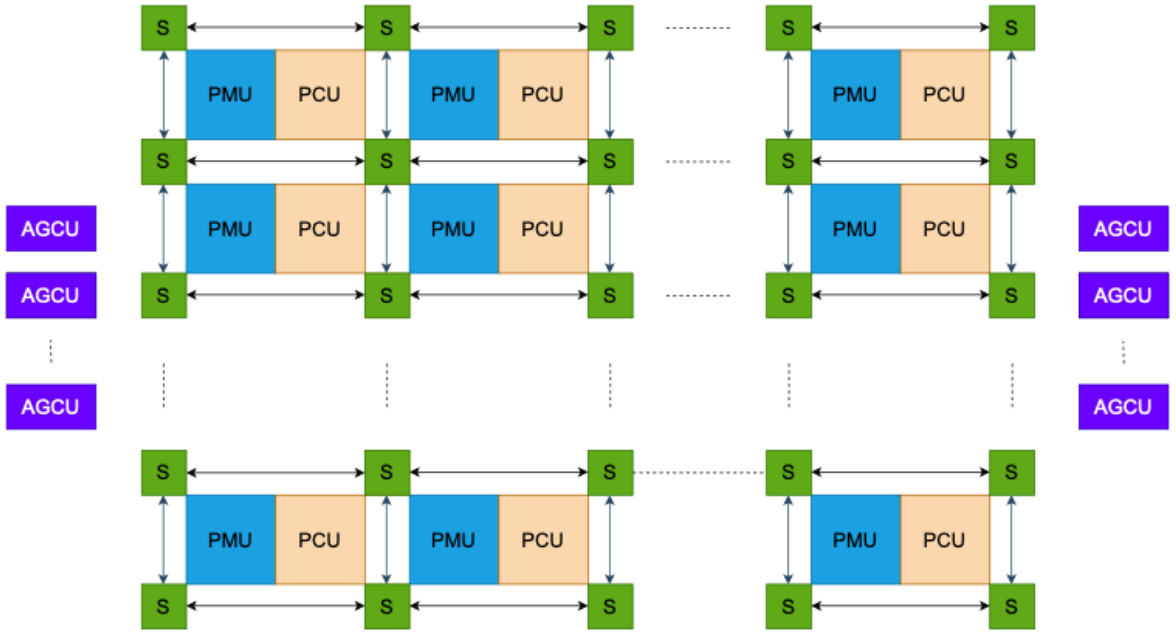
Low Latency Model Switching

(E.g., < 0.01s for llama3.1 8B)

SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

SN40L Chip: Tile Architecture



1040 PCUs and PMUs

PCU: Compute unit

PMU: Memory unit

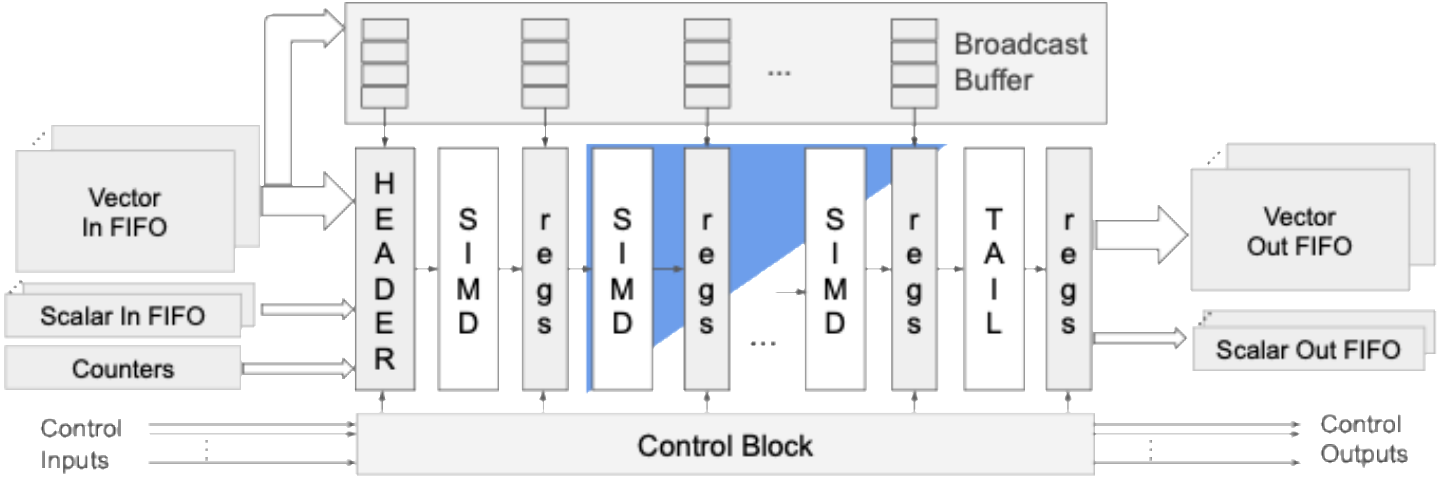
S: Mesh switches

AGCU: Portal to off-chip memory and IO

SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

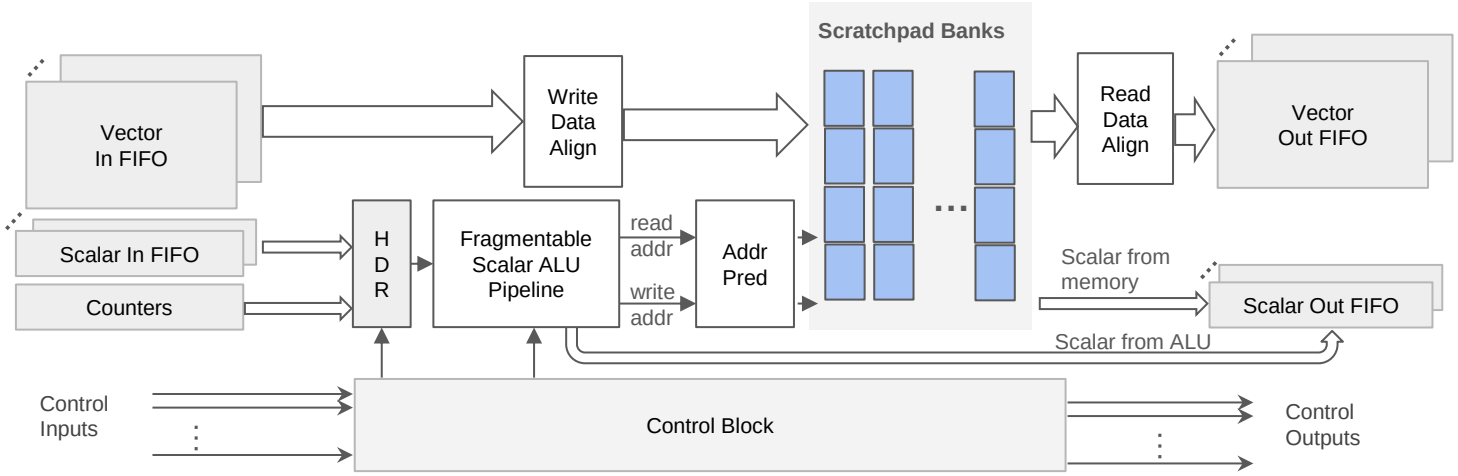
SN40L PCU



- Configurable as a systolic array or a SIMD vector unit with M lanes
- BF16, FP32, INT32, and INT8 compute data types, configurable storage data types
- Arithmetic, Logical, and Bitwise operations
- A cross-lane reduction tree (blue) to reduce along the vectorized dimension
- Tail stage provides transcendental functions, casting, and stochastic rounding capabilities

Copyright © 2024 SambaNova Systems Inc.

SN40L PMU



- Programmer managed scratchpad memory supports concurrent reads and writes
- Fragmentable, address-generation pipeline that can produce 4 addresses per cycle
- Data alignment crossbars enable high throughput tensor transformations such as transpose, dilation, downcast
- Address predication support enables composing multiple PMUs to store a large tensor

Copyright © 2024 SambaNova Systems Inc.

SN40L Unit-Level Interconnect

- Mesh Interconnect made of 3 physical networks:
  - **Vector**: Carries data, Packet Switched
  - **Scalar**: Carries addresses, other metadata, Packet Switched
  - **Control**: Carries flow control, synchronization and graph orchestration tokens
- Point to Point, One-Many Multicasts, and Many-One transmissions
- Hardware 2-D dimension order routing (DOR) with software override
- **Flow control** with packet-level E2E credits, or Tensor-level software control tokens

Copyright © 2024 SambaNova Systems Inc.

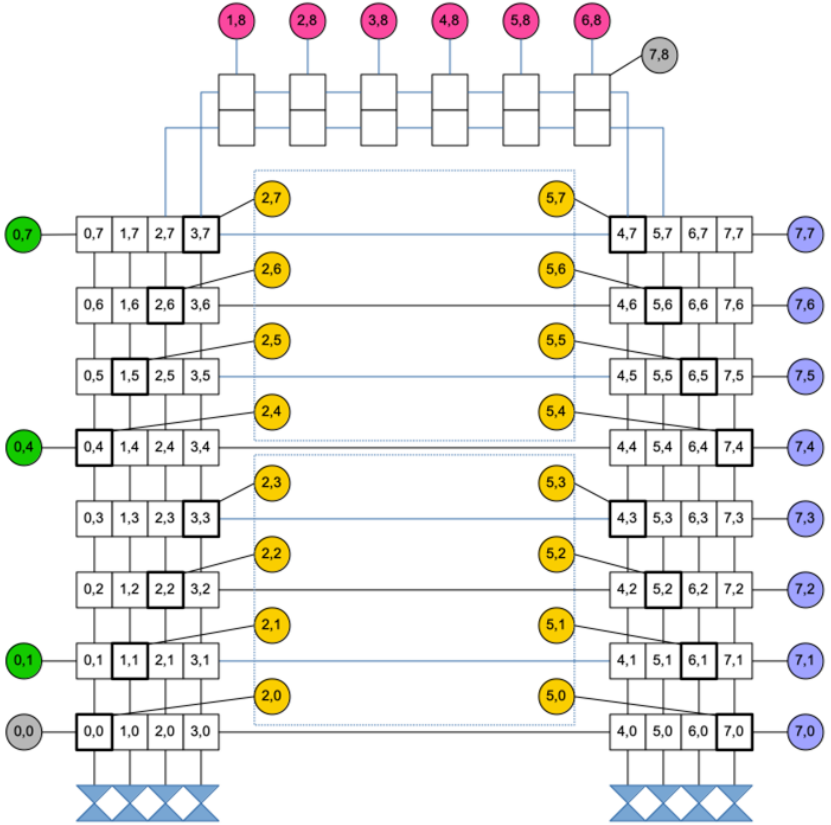
SN40L AGCU

- **Memory address generation** to access DDR, HBM, Host
- **Peer-to-Peer communication** between RDUs for collective communications
- **Graph control interface** to load and orchestrate graph execution
  - Enables RDU to orchestrate a schedule of graphs without host involvement
- **Segment Lookaside Buffer** for virtual-physical address translation and memory access management

Copyright © 2024 SambaNova Systems Inc.

SN40L Top-Level Interconnect

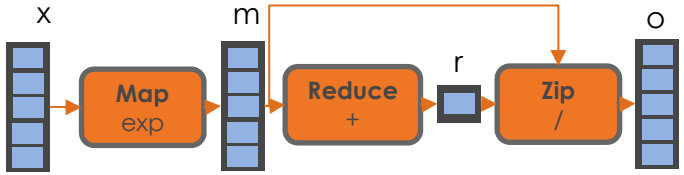
- Hybrid Mesh/Ring Packet Switched Interconnect with four physical networks:
  - **Data**: Read and write data
  - **Request**: read and write requests
  - **Response**: Write acknowledgement and interrupts
  - **Credit**: End to end credit exchange
- Routing of packets uses a Y-X DOR mechanism
- All traffic is managed by end-to-end credits or guaranteed consumption to avoid deadlocks

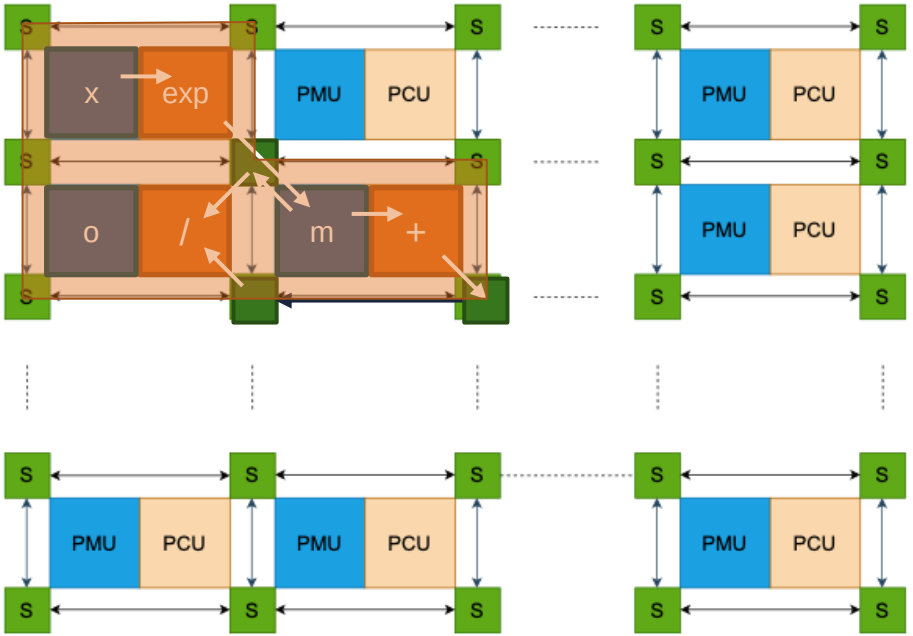


Copyright © 2024 SambaNova Systems Inc.

Example: Softmax

SOFTMAX: $\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$

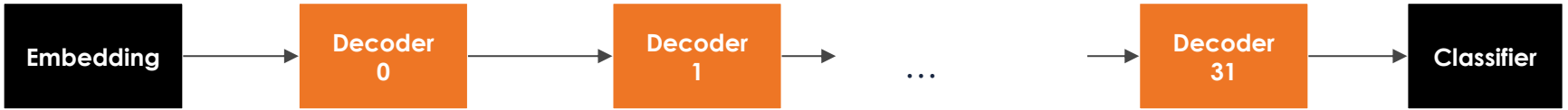




Copyright © 2024 SambaNova Systems Inc.

Structure of a Transformer

Example: Llama3.1 8B

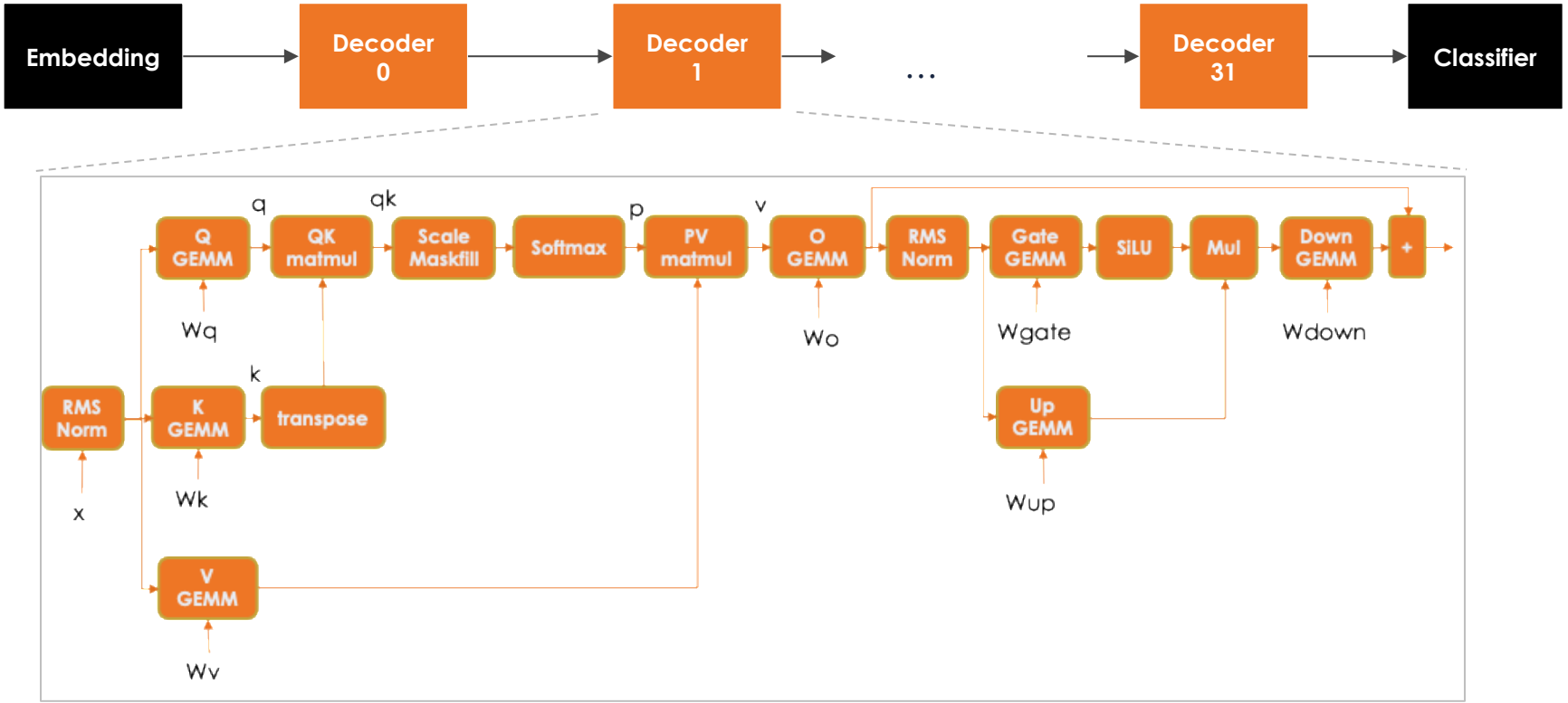


SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

Structure of a Transformer

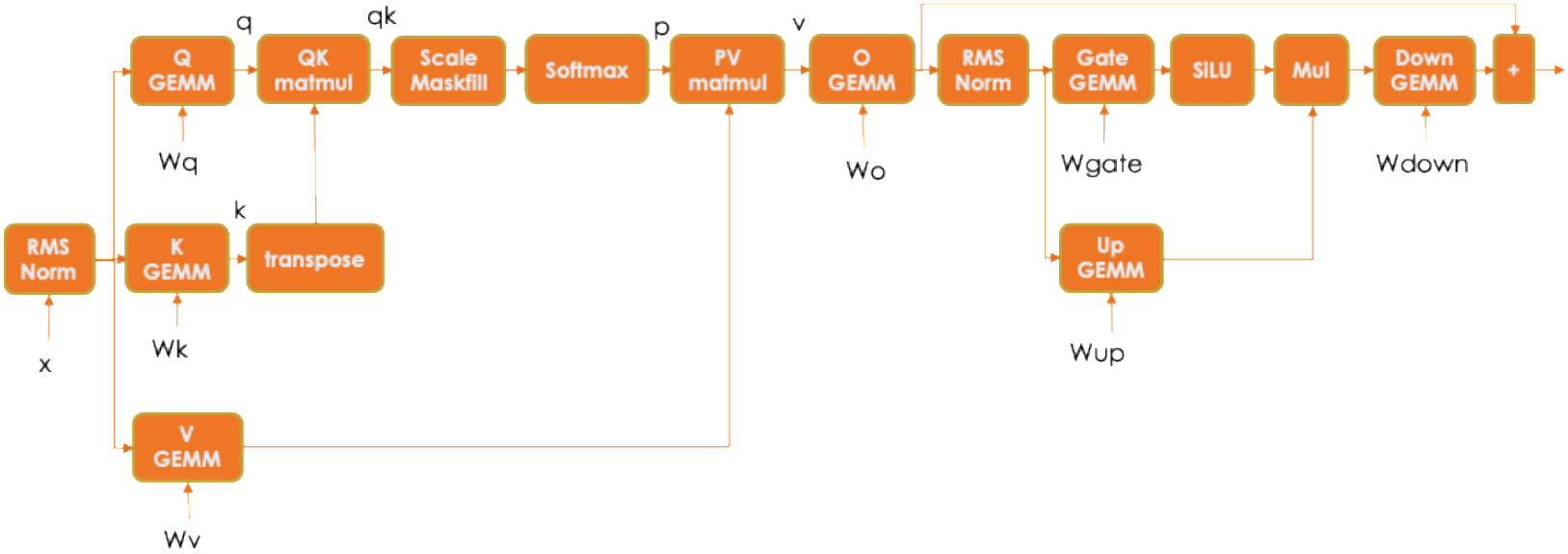
Example: Llama3.1 8B



Copyright © 2024 SambaNova Systems Inc.

Decoder

Example: Llama3.1 8B

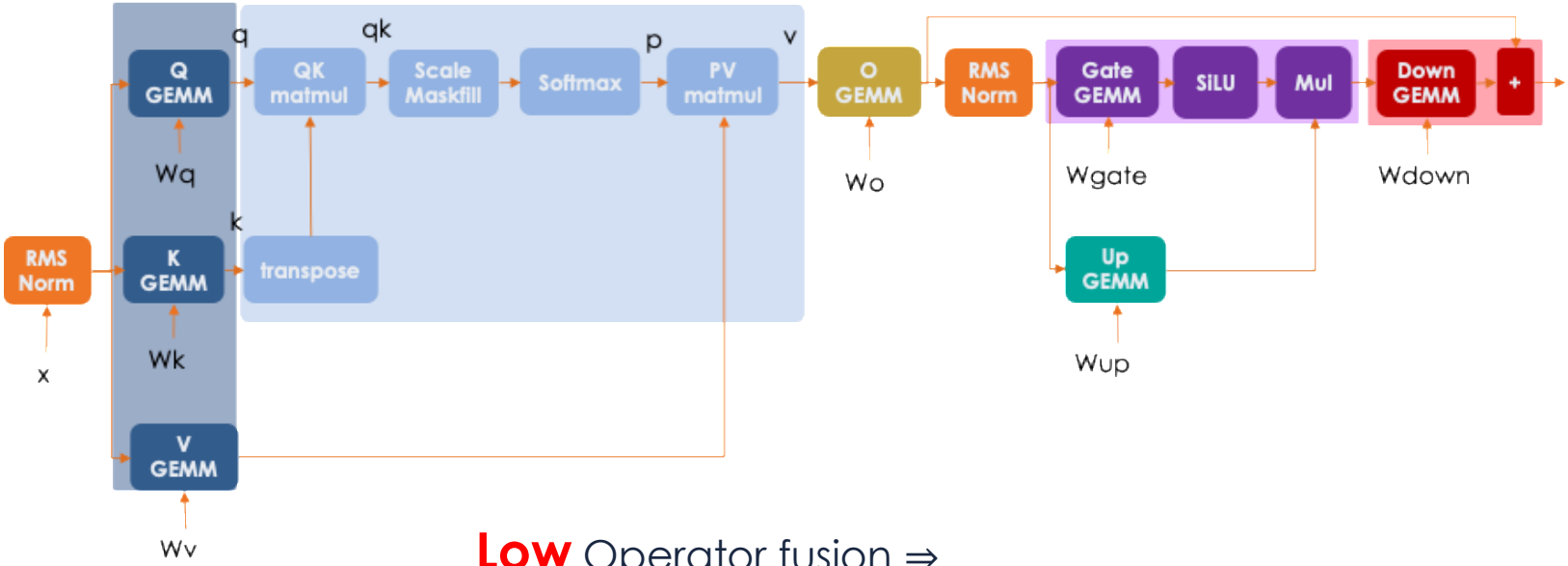


Copyright © 2024 SambaNova Systems Inc.

Decoder: Conventional Fusion (GPU)

Example: Llama3.1 8B

1 colored group == Fused ops == 1 kernel call



Low Operator fusion ⇒

High Kernel Launch Overheads

Low data locality

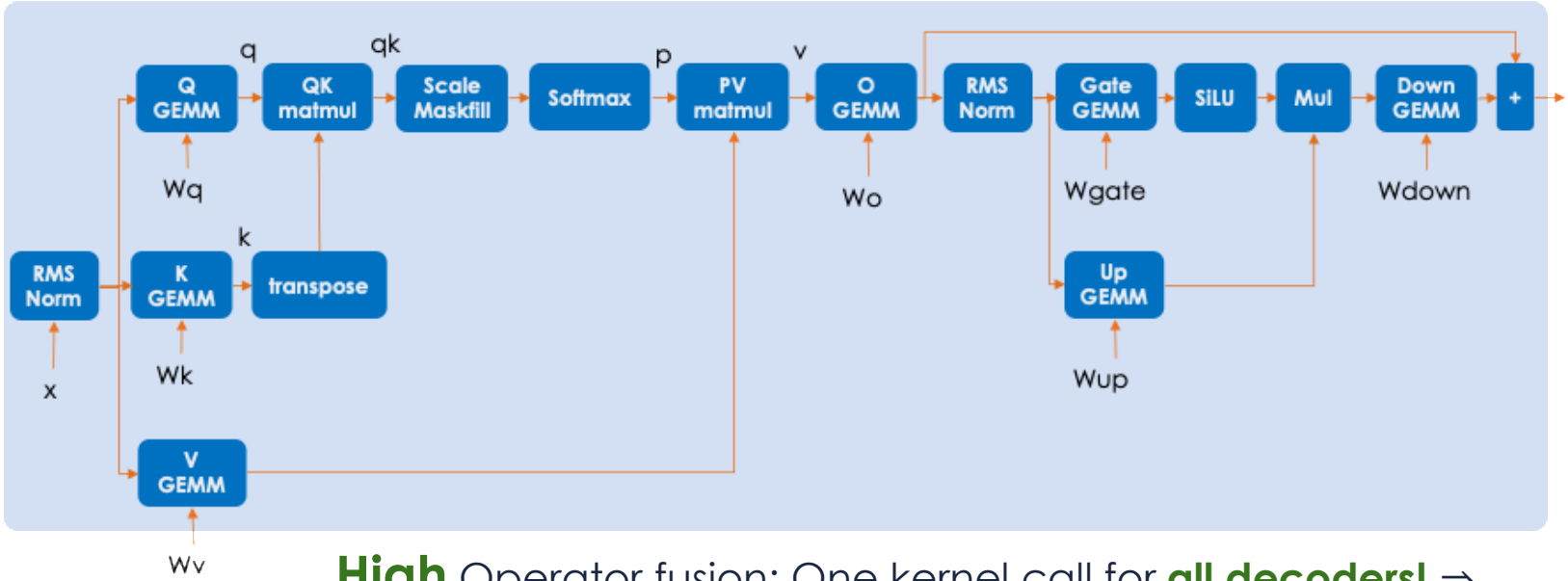
SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

Decoder: Streaming Dataflow (RDU)

Example: Llama3.1 8B

1 colored group == Fused ops == 1 kernel call



High Operator fusion: One kernel call for **all decoders!** ⇒

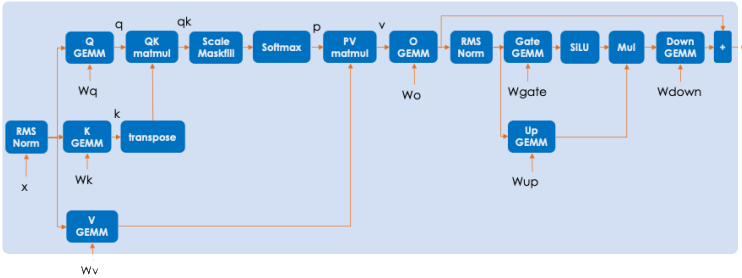
Zero Kernel Launch Overheads

High data locality

SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

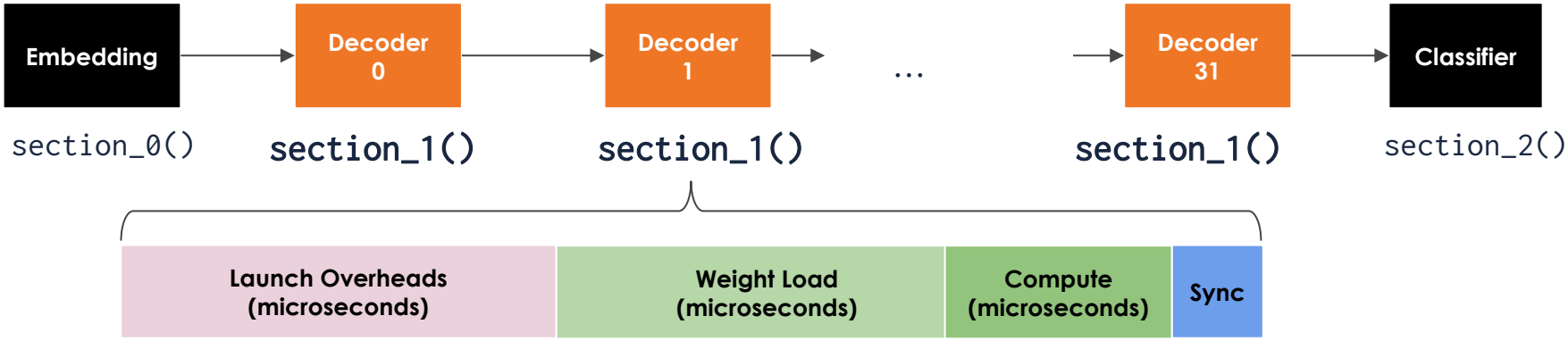
Decoder as a single kernel on RDU

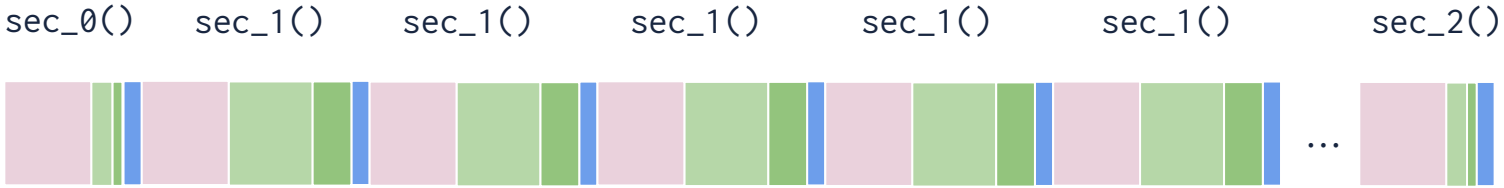


Copyright © 2024 SambaNova Systems Inc.

Executing Transformer on RDU

Example: Llama3.1 8B

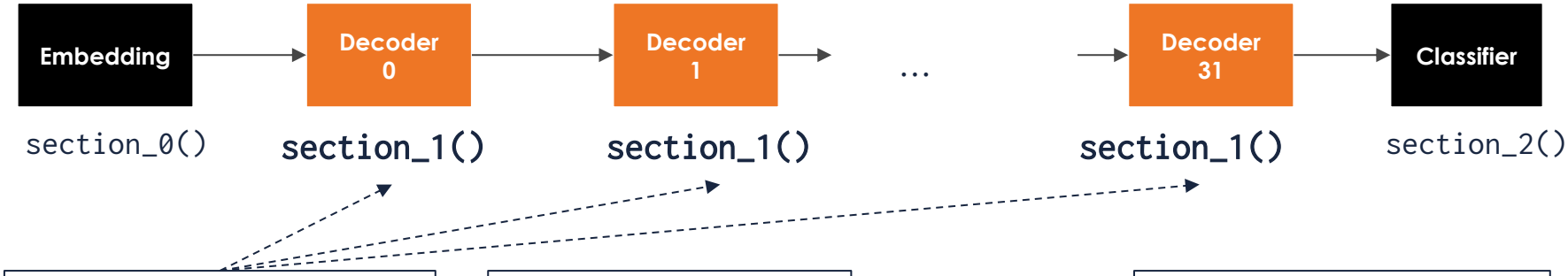




SambaNova SYSTEMS
Copyright © 2024 SambaNova Systems Inc.

RDU Hardware Graph Orchestration: Temporal Fusion

Example: Llama3.1 8B



Single Section Decoder
+
Repeated Section Calls
=
New fusion opportunities!

```
section_1(arg):
  Foo;
  Bar;

// Repeat 32 times
section_1(arg0)
section_1(arg1)
…
section_1(arg31)
```

LoopConversion

```
section_1(args[]):
  for i in range(32):
    Foo;
    Bar;

// 1 time
section_1(args)
```

1 kernel launch for all decoders!

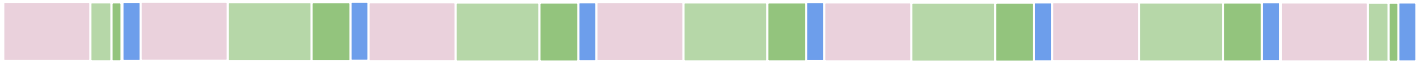
Copyright © 2024 SambaNova Systems Inc.

Executing Transformer on RDU

Example: Llama3.1 8B

Baseline: **100** tokens/s

+ Single Section Decoder



= **500** tokens/s
16 SN40L

+ Loop Conversion

= **1100** tokens/s
16 SN40L

Copyright © 2024 SambaNova Systems Inc.

405B, 70B and 7B llama3.1 on 16 SN40L Chips

Highest tokens/s with full 16-bit precision

Model	Weights Precision (bits)	Activation Precision (bits)	Tokens/ second
7B llama3.1	BF16	Mixed (BF16, FP32)	**1100**
70B llama3.1	BF16	Mixed (BF16, FP32)	**380**
405B llama3.1	BF16	Mixed (BF16, FP32)	**114**

Try It Yourself!

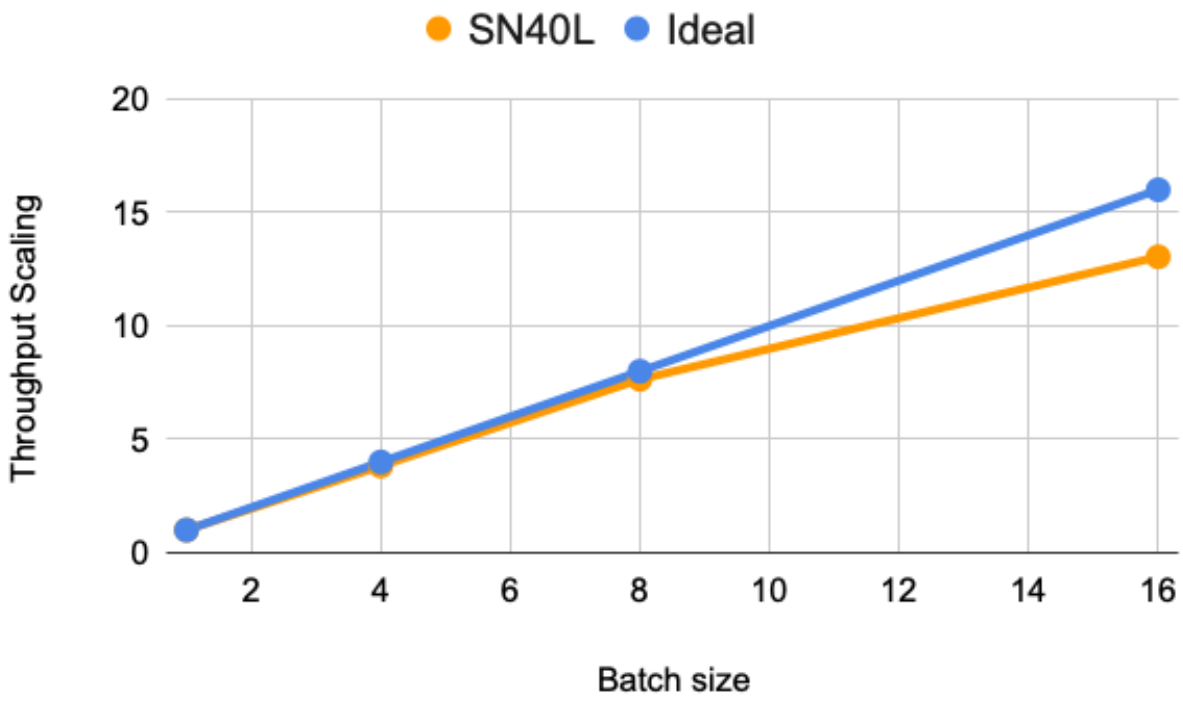
https://fast.snova.ai/

SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

Batched Inference and Throughput Scaling with Llama 70B

Close to Ideal Throughput Scaling with Batch Size on Llama 3.1 70B



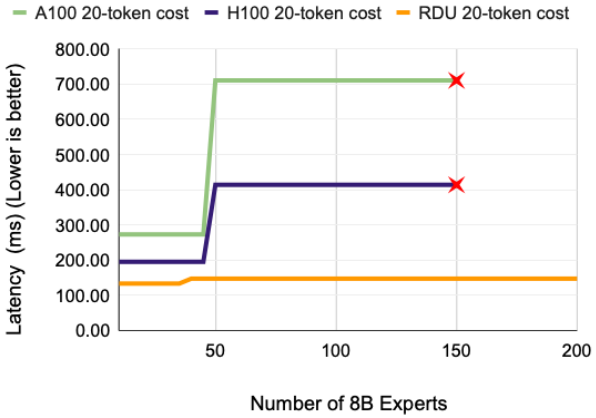
Try It Yourself!

https://fast.snova.ai/

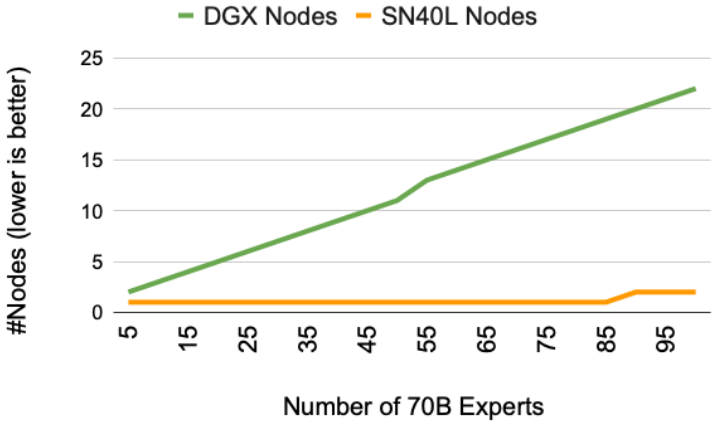
Copyright © 2024 SambaNova Systems Inc.

Much Lower Model Hosting and Switching Costs

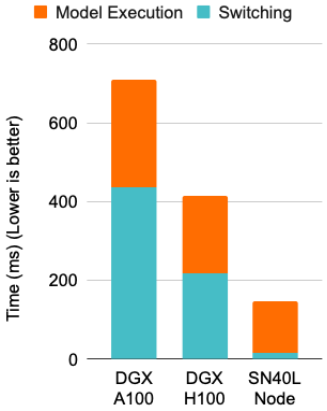
Host and Serve Trillions of Parameters at a fraction of the cost of GPUs



3.7x Faster
vs. DGX H100



19x Smaller
Machine Footprint



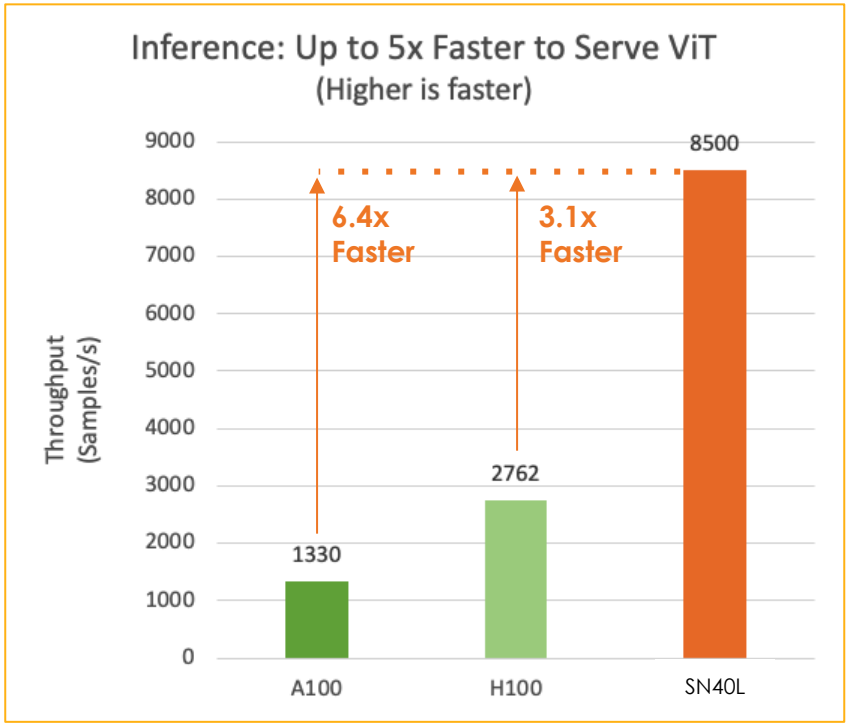
15x Faster
Model Switching

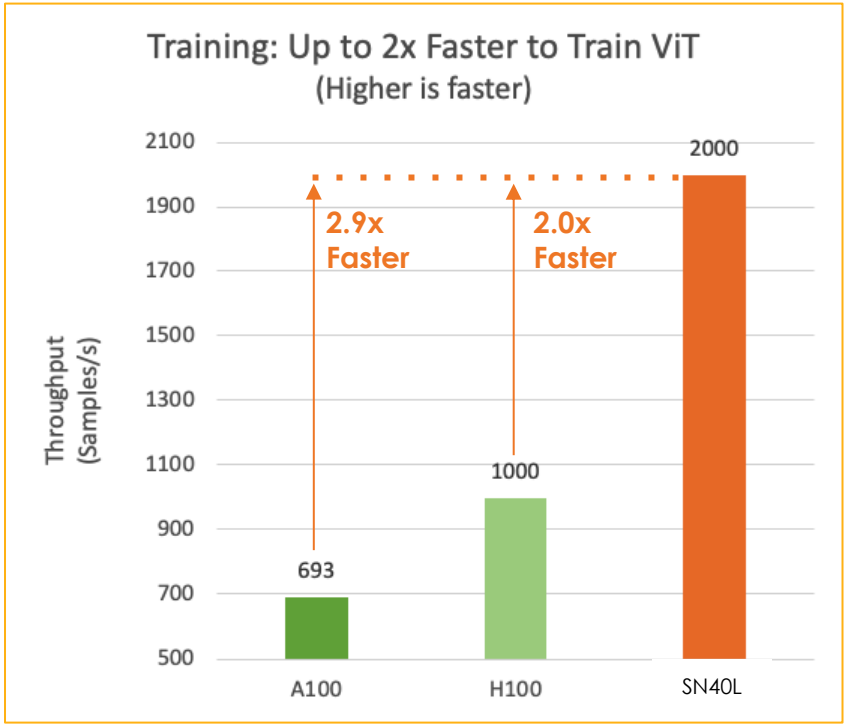
SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

Fast Tracking Multimodality Implementations

Up to 6.4x faster to infer and 3x faster to train (ViT)





* These numbers are measured on ViT-Based with 257 classes using patch size of 32

SambaNova SYSTEMS

Copyright © 2024 SambaNova Systems Inc.

Try It !

https://fast.snova.ai

API Access (free!)

https://sambanova.ai/fast-api

Copyright © 2024 SambaNova Systems Inc.