Evolving Inference Processor Software Stacks to Support LLMs

Ramteja Tadishetti

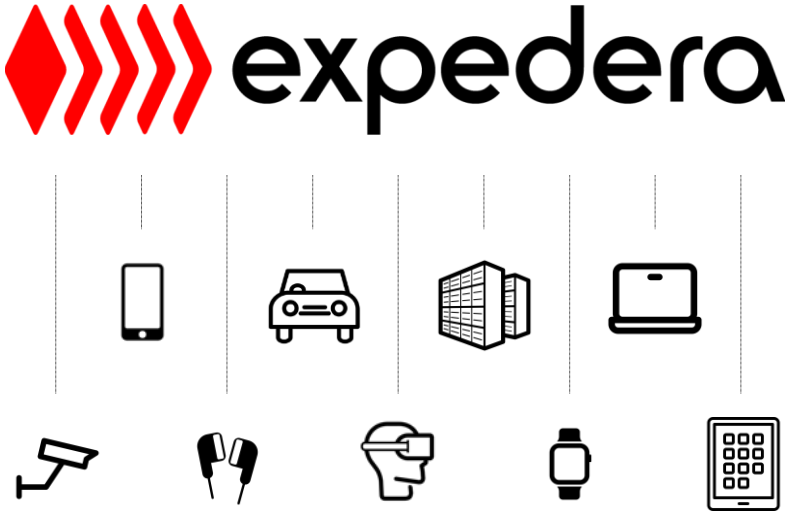
Principal Software Engineer

Expedera Inc.

expedera

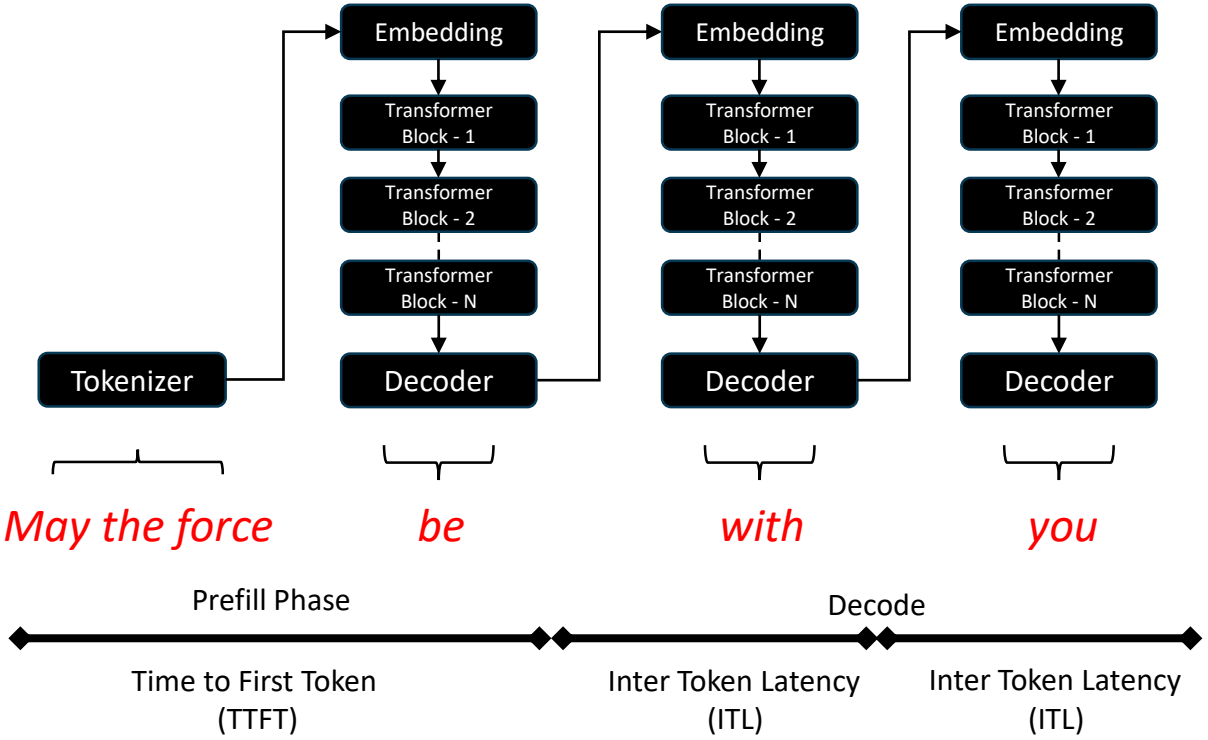
Intro to Expedera

- Silicon Valley startup company offering optimum artificial intelligence NPU inference IP based on revolutionary packet processing architecture
- Use case-customized with native support for hundreds of LLMs, CNNs and others
- Multiple production customers in smartphone, automotive, consumer device and data center markets
- Field- and silicon-proven: well over 10M devices shipped worldwide with Expedera IP



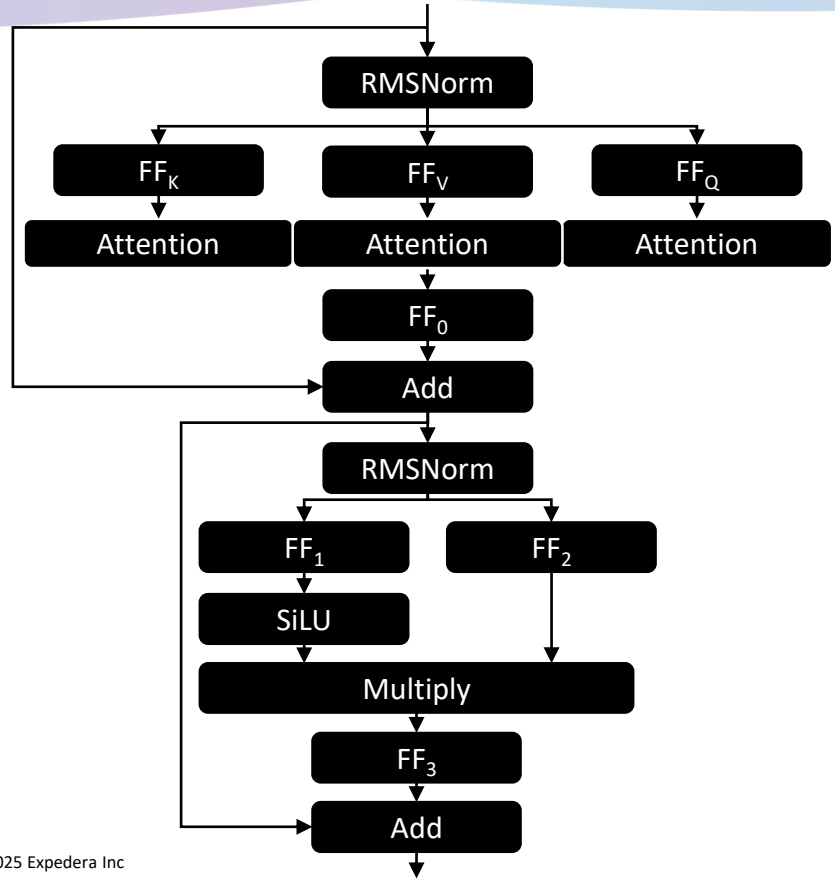
LLM Inference Flow

- User specifies input prompt
- Tokenizer converts prompt into tokens
  1. All Input tokens can be processed simultaneously (prefill phase)
  2. Generated output token is used as input in next iteration. (decode phase)



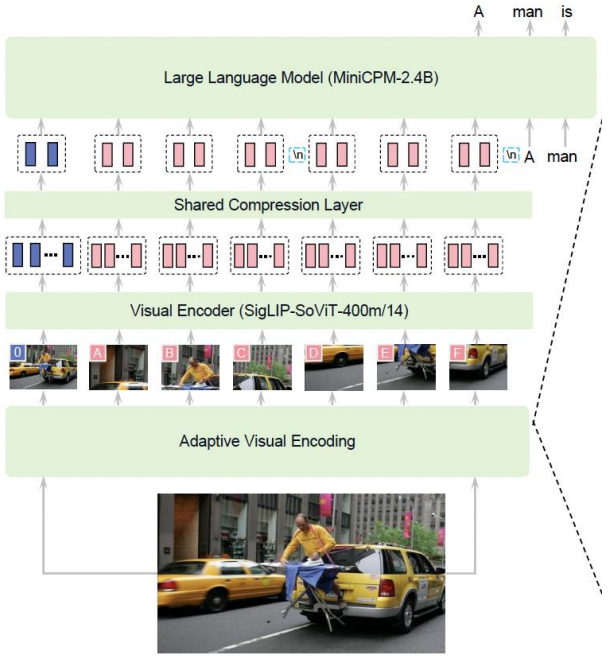
Llama 3.2 1B Transformer Block Architecture

- 1 billion parameters
- 2 billion operations without attention
- Operations in attention increase with context size (n)
  - $O(n^2)$ parallel compute during prefill
  - O(n) * n sequential compute during decode per token



LLMs in the Vision Domain

- MiniCPM-V: multimodal LLM, built on SigLip-400M and Qwen2-7B
  - Introduces new features for multi-image and video understanding
- SigLip (ViT-based) model is used to extract vision embeddings for each image
  - Video is converted to a series of images
- Resampler used to compress vision embeddings to 96 tokens per image
- Qwen2 LLM processes prompts (text + vision) and generates text



"*MiniCPM-V: A GPT-4V Level MLLM on Your Phone*" Yao et all, 2024

Increased Runtime Inference Complexity with LLMs

CNN Network

LLM (Multi-turn)

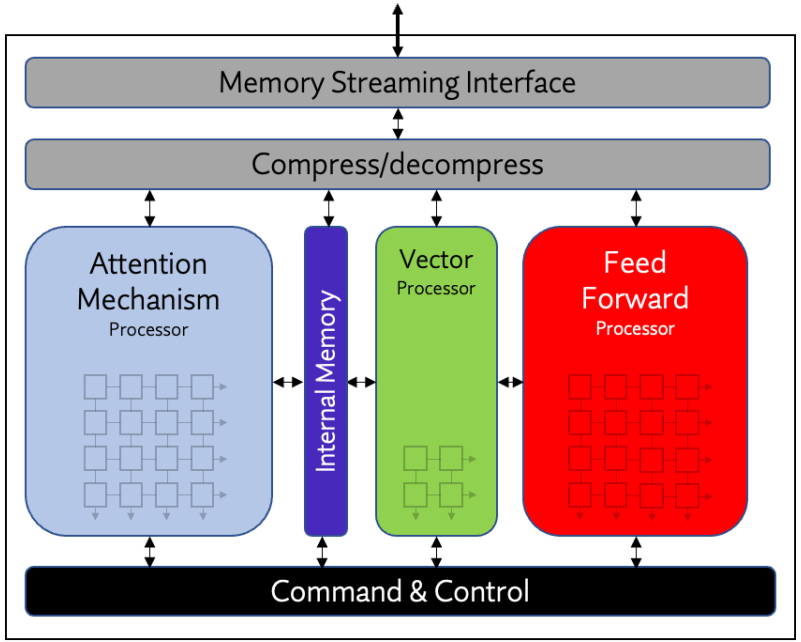
- Prefill tokens can contain user history, generation directive or additional retrieved knowledge or older retired conversation
- Decode phase streams tokens and generate responses
- Sessions are inactive until next user input is received
- Follow-up prefill processes user input specified during inactivity

LLM Inference Solution Considerations

Application Needs	Compilation Challenges	Runtime Requirements	Hardware Design
• Accuracy • Latency • Context Size • Quantization & mixed-precision requirements	• Representation • Dynamic shapes	• Multi-turn • Multi-modality • Speculation	• Cost • Performance • Bandwidth

Expedera Origin Evolution™ Packet-based NPU AI Engine

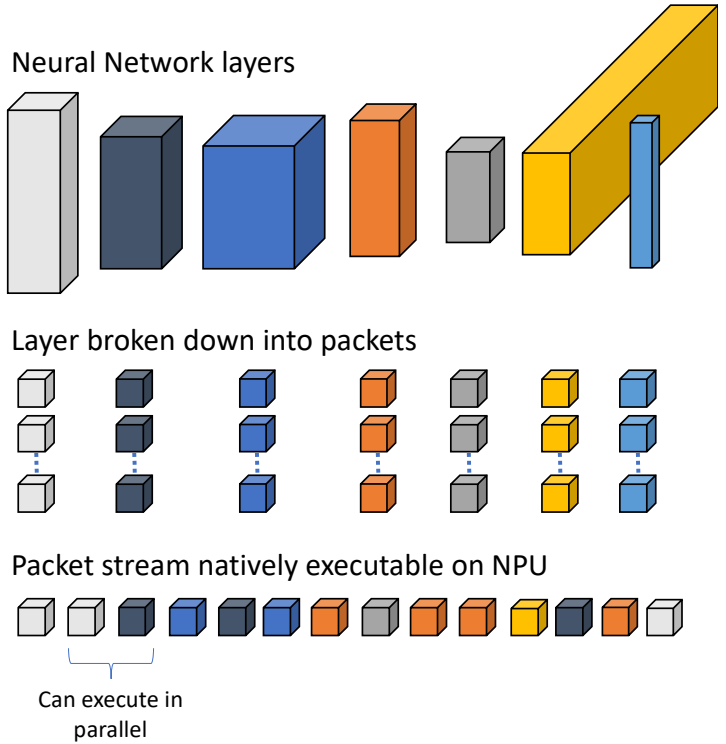
- Revolutionary packet-based architecture
- Deeply connected pipelined approach
  - Packetization
  - Weight streaming
  - Optimized runtime APIs for LLM inferencing
- Processes models as-is; no re-training or accuracy degradation required
- Single engine performance to 128 TFLOPS, scalable to PetaFLOPS
- In customer silicon now



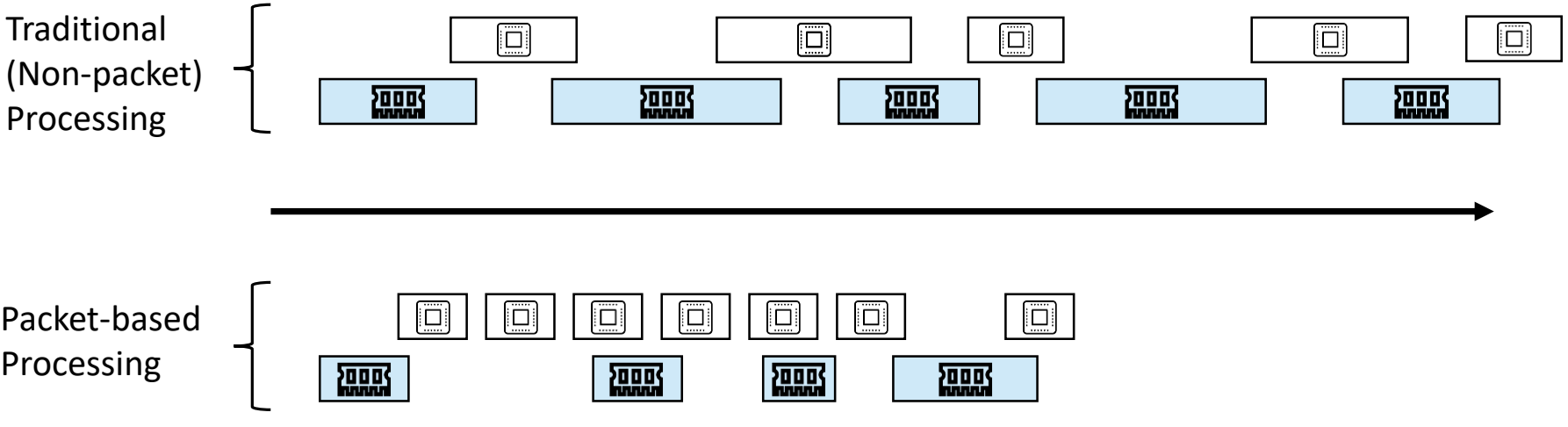
Origin Evolution Architecture

Packets – a Unique Approach to AI Optimization

- Packet - a contiguous fragment of a neural network layer with the entire context of execution
- Manages activations intelligently by inferring fine-grain dependencies
- Minimal data movements
- Greatly increases performance while lowering power, area and bandwidth requirements

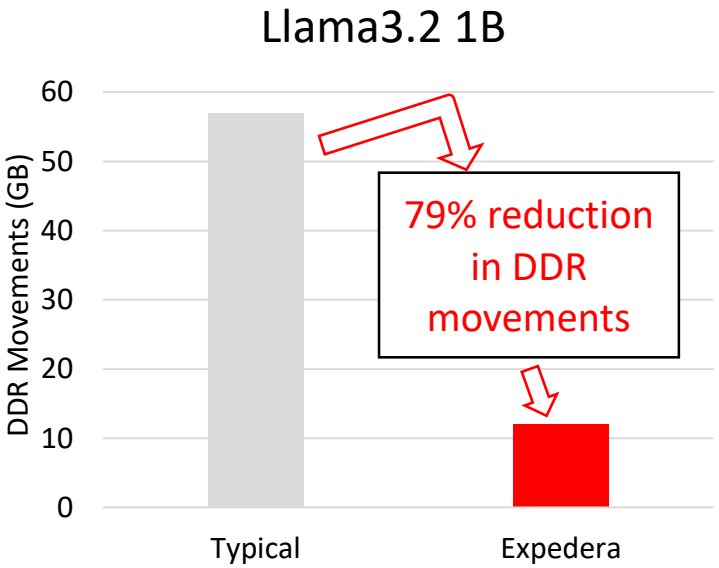


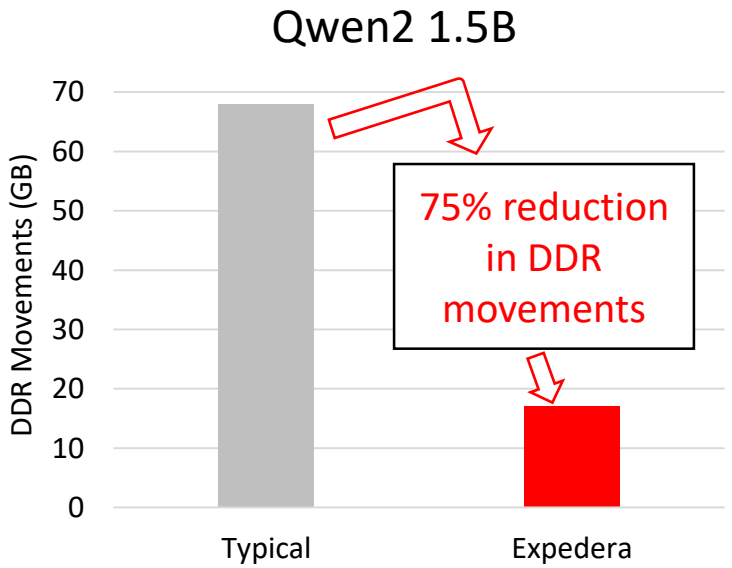
Inference Event Timeline Comparison



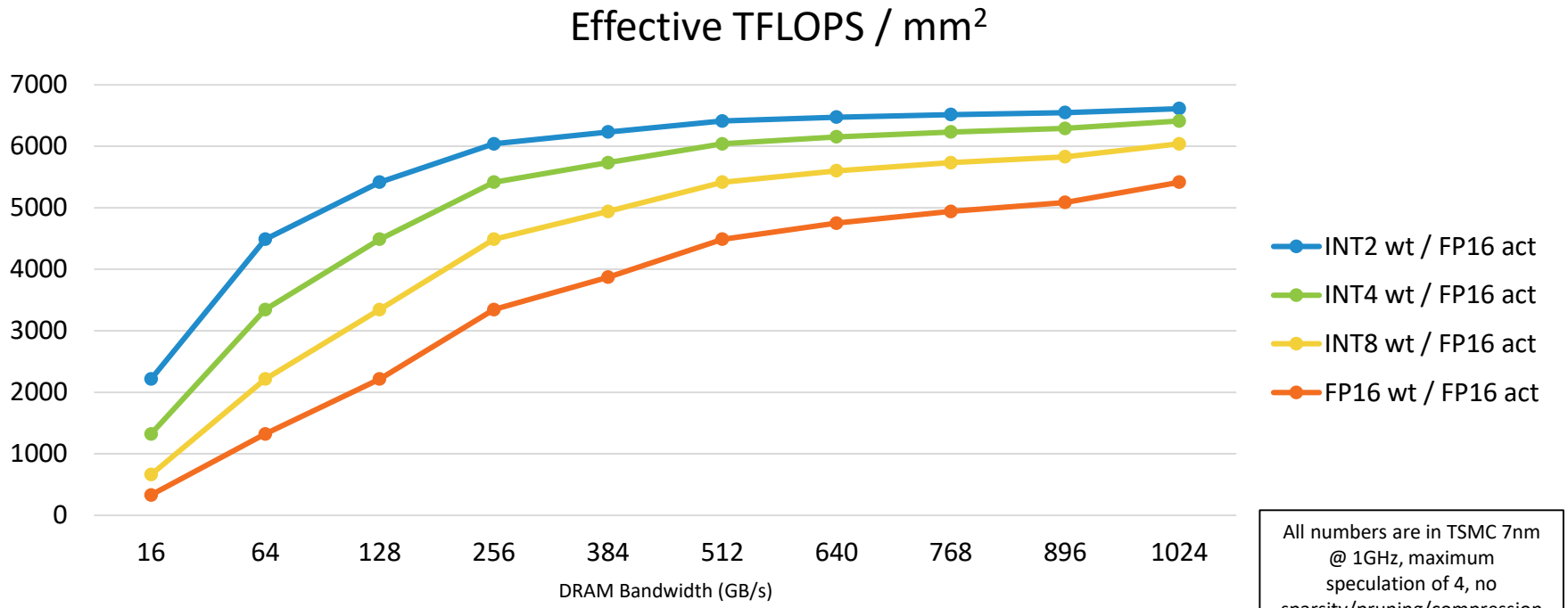
More efficient use of compute resources (higher utilization) with fewer memory moves (lower power consumption)

Reduced DDR Movements with Packetized Architecture





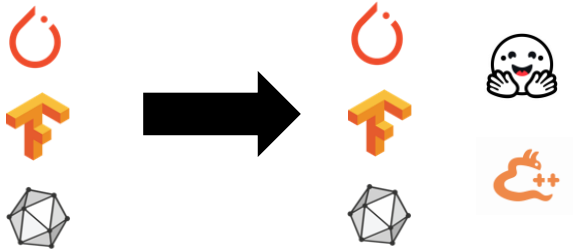
Highest Effective TFLOPS



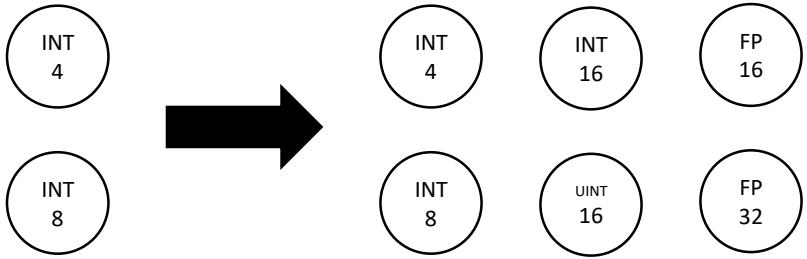
All numbers are in TSMC 7nm @ 1GHz, maximum speculation of 4, no sparsity/pruning/compression applied (though supported)

Software Stack Evolution Considerations

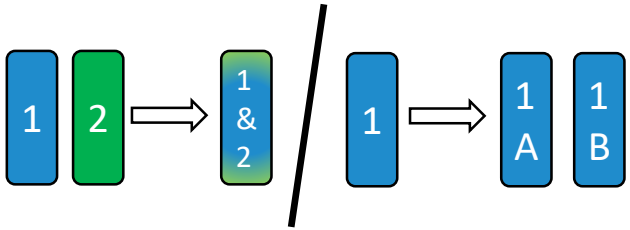
- Representations



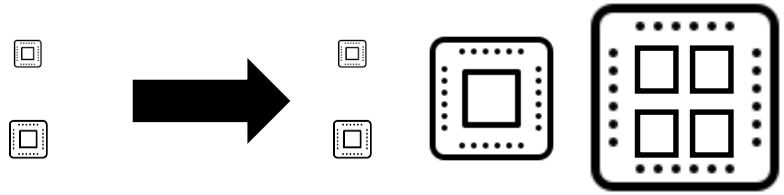
- Precisions (including Mixed Precision)



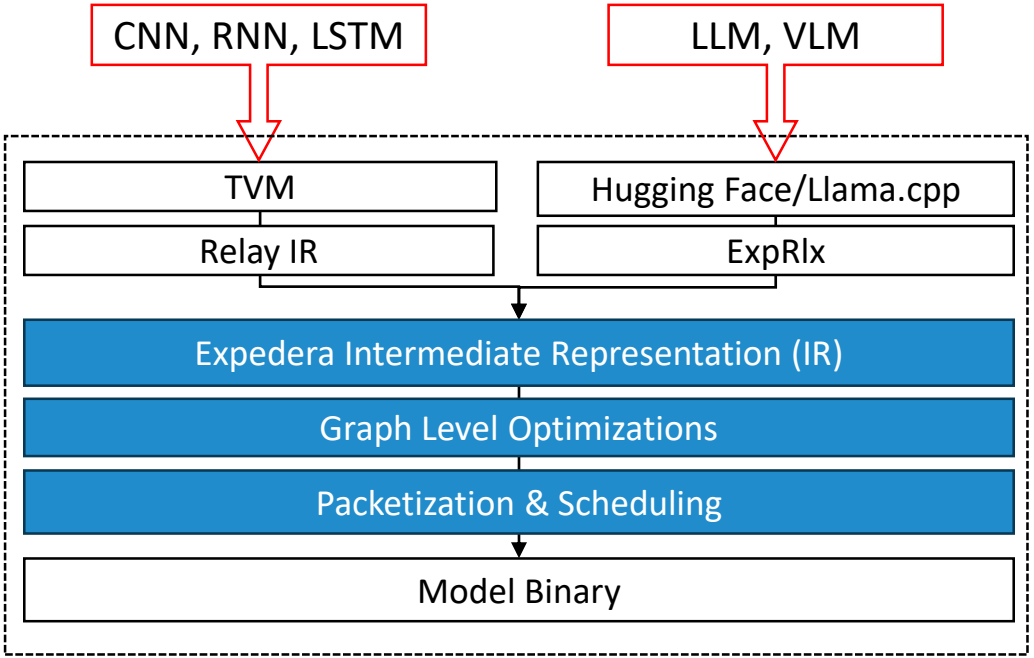
- Layer Fusions and Fissions



- Scaling TOPS/TFLOPs and # of Cores



Expedera Origin Evolution LLM Software Stack



In Conclusion

- LLMs present a unique set of difficulties to inference hardware and software stacks: model sizes, runtime operations, transformers, and solutions considerations
- Software stacks must evolve to support more representations, more varied precision types, layer fusion/fissions, and to support higher TOPS/TFLOPS and multi-core implementations
- Packetization architecture enhances utilization and reduces memory movements, essential for LLMs
- Expedera's Origin Evolution is available today to support LLM inference needs

Resource Slide

Summit & Alliance Resources

- Visit us at booth #520
- Alliance website
  - https://www.edge-ai-vision.com/companies/expedera/

Expedera Resources

- Company Website
  - http://www.expedera.com/
  - White papers, technical briefs, webinars, other
- Pre-silicon PPA Estimates
  - Want cycle-accurate PPA numbers for your use case(s) well before silicon?
  - info@expedera.com
- Contact us directly
  - info@expedera.com