AI Tokenomics Explained

Shruti Koparkar *Sr. Manager, Product Marketing, NVIDIA*

Eduardo Alvarez *Manager, Technical Marketing Engineering, NVIDIA*

Two Key Levers Impacting Economics of AI

What tokens you choose?

- Model complexity
- Context length
- Interactivity



What infrastructure you choose?

- Infrastructure cost
- Throughput per Watt
- Token Cost



What Tokens You Choose: Not All Tokens Are Created Equal



Model Size and Complexity

Tokens produced by more complex models carry more intelligence



Long Context

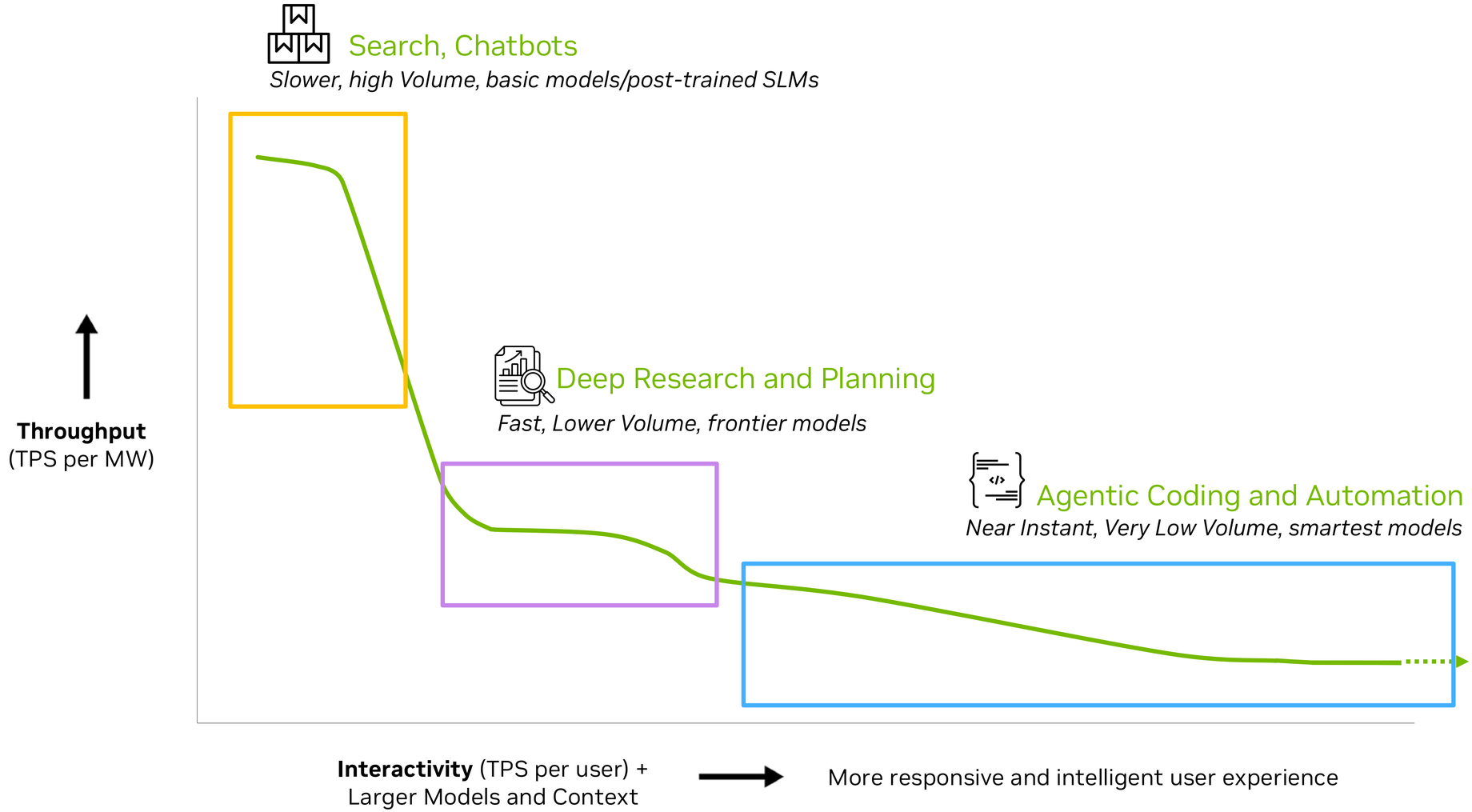
Tokens produced with longer context carry more intelligence



High interactivity

Tokens that arrive faster are more valuable

Map Usecases To Optimal Token Value



What It Looks Like In Practice?



Right model for right task

Use premium models when necessary



Real-time routing

Intelligent routing based on model fit, pricing, and context



Governance and observability

Build a control plane that provides visibility into token use

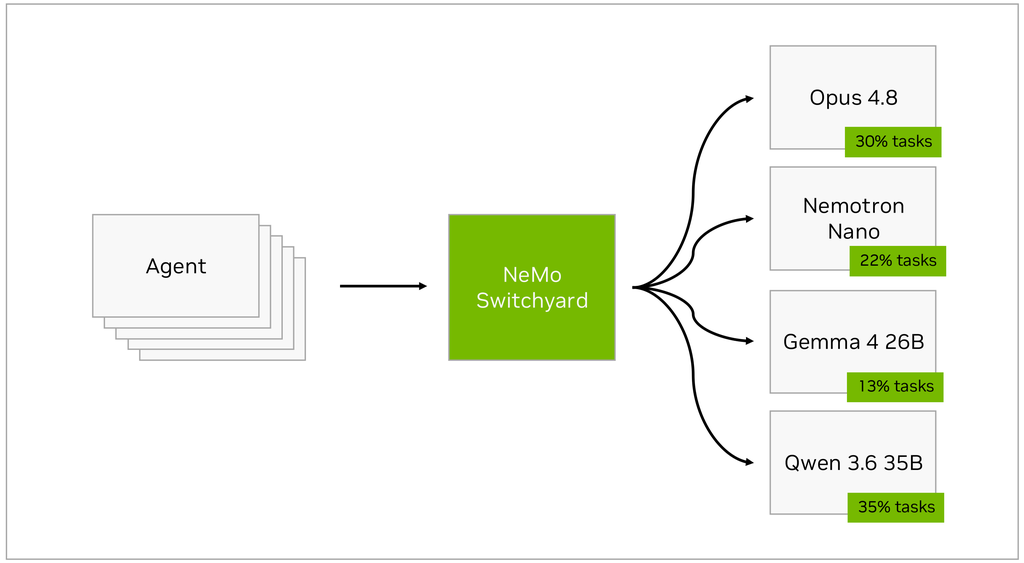
NVIDIA NeMo SwitchYard: 3X Lower Cost Per Agentic Task

Open source model routing library for AI agents helps map agent needs to model capabilities

Configurable family of routing algorithms

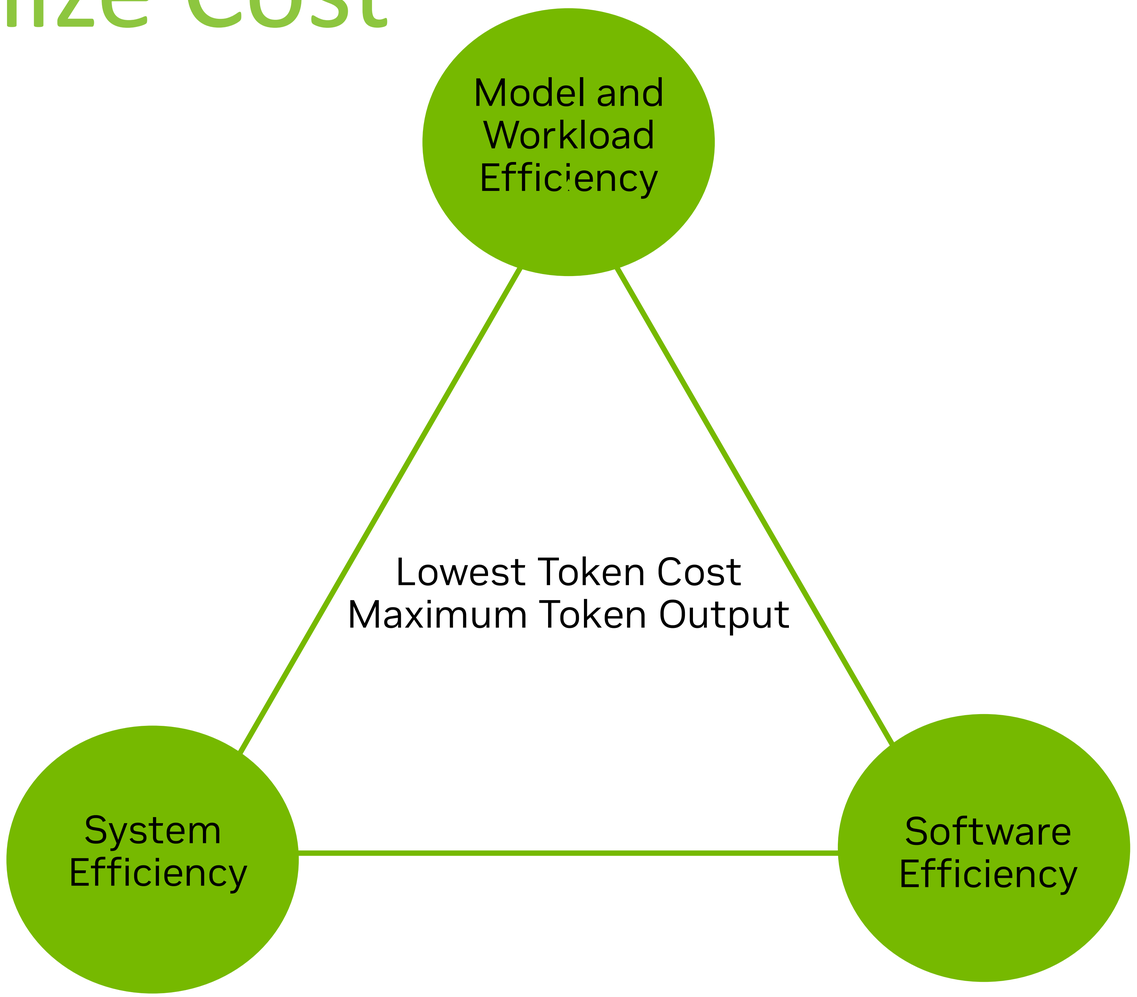
Works with open and closed models

Compatible with leading model APIs, agents, SDKs and frameworks



	Single model	NeMo Switchyard
LLMs	1	4
Tasks completed	89	90
Total cost	$179	**$60**

What Infrastructure You Choose: Maximize Tokens, Minimize Cost



Model Efficiency: Mixture of Experts

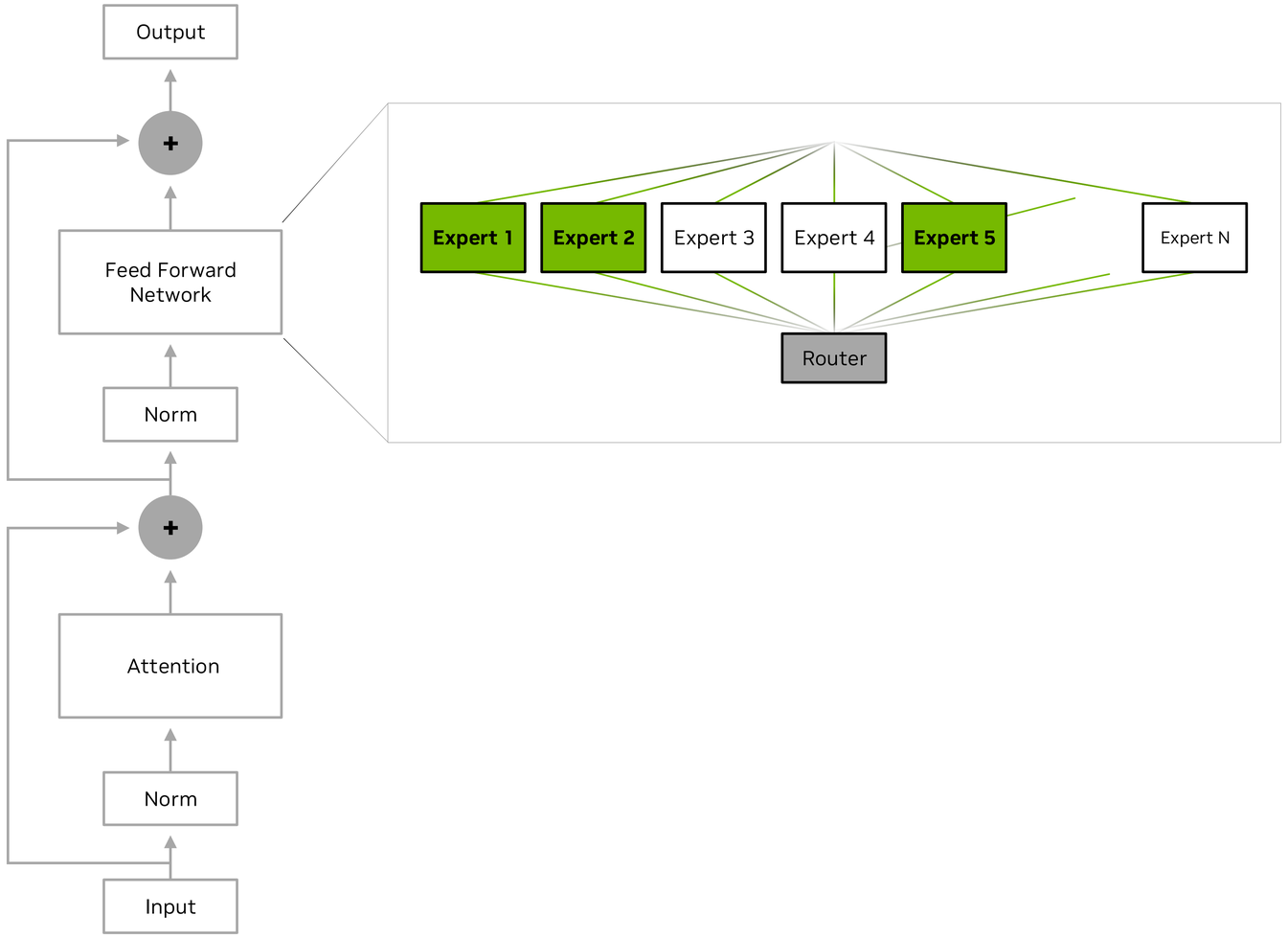
Kimi K3

2.8 Trillion parameters, 108 Billion active

896 experts / 16 activated per token

Lower per-token compute and
HBM bandwidth

But...

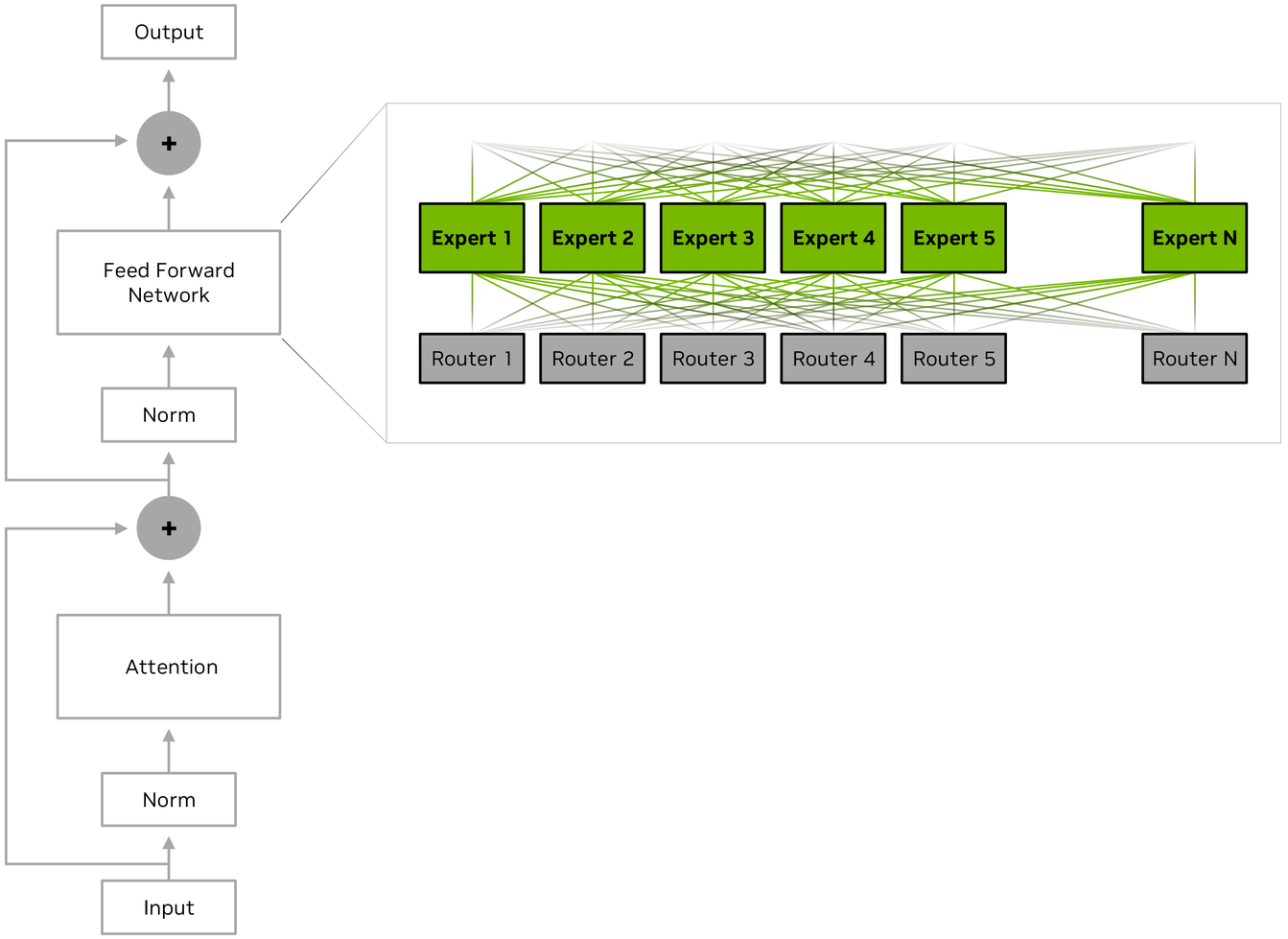


Communication is Hidden Cost of Mixture of Experts

But…

Heavy all-to-all traffic for multiple tokens across multiple batches

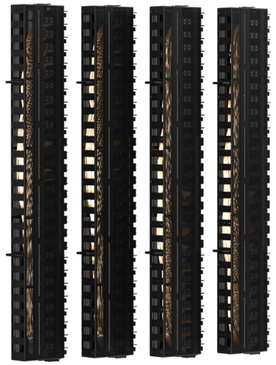
Strong dependence on scale-up network bandwidth



System Efficiency: Fast Inter-GPU Communication with Purpose-built Scale-up Networking



NVL72 Rack
72 GPU scale-up domain



NVLink Spine



NVLink 6 Switch Tray

3x
Lower Latency

10x
Higher Packet Rate

130 TFLOPS
In-Network Compute

Agentic AI Is Significantly More Demanding

1K → **142K**

ISL (median)

Kernels for broader set of shapes/sizes

Disaggregated serving with rate matching

60% → **98%**

Cached Tokens

Prefix caching

Distributed KV caching

KV-cache offload

KV-aware routing

3 → **65**

Number of Turns

Session-aware prefix-cache retention

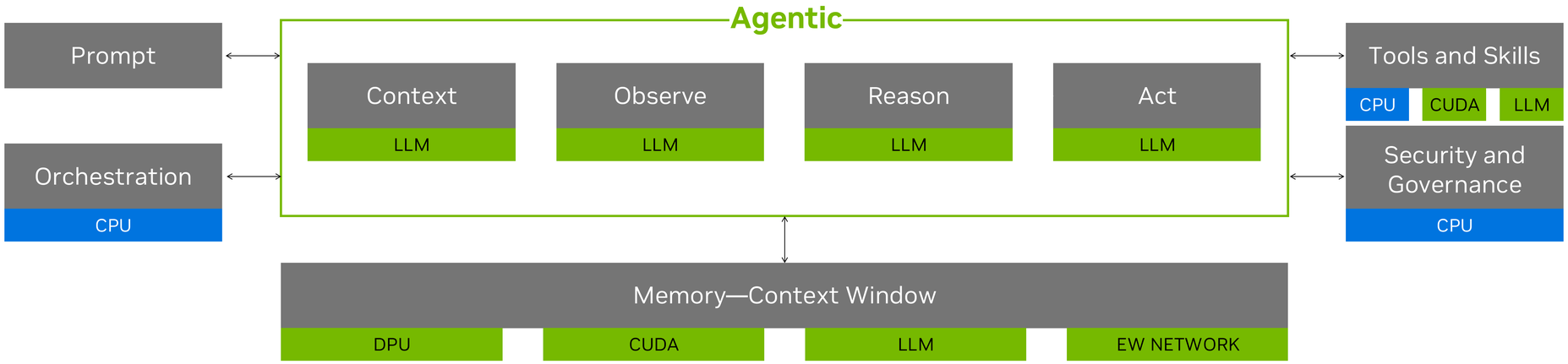
Interval-based and selective checkpointing

0 → **4**

Sub-Agents

Intelligent scheduling and routing

KV cache needs to be preserved for longer



Source: SemiAnalysis AgentX workload characteristics

Software Efficiency: Enable Every Optimization

Serving and Orchestration

Distributed Serving		Orchestration		Memory Management	
Dynamo	llm-d	Grove/K8s	Run.ai	KVBM	LMCache

Application Acceleration

Runtime		
SGLang	TensorRT-LLM	vLLM

Accelerated Kernels		CUDA Libraries			Model Compression
CUTLASS	FlashInfer	cuDNN	cuBLAS	cuDF	ModelOpt

KV Cache Transfer		Collective Communication
NIXL	Mooncake	NCCL

Infrastructure Access

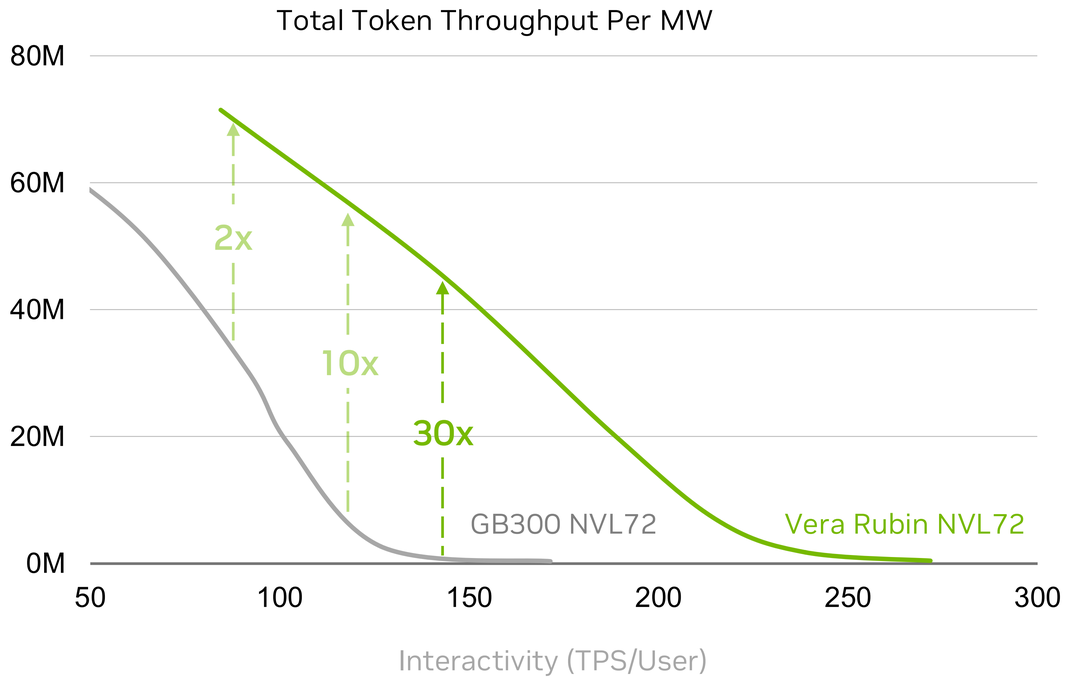
GPU	DPU	Storage	
CUDA	DOCA	DOCA Memos	NIXL

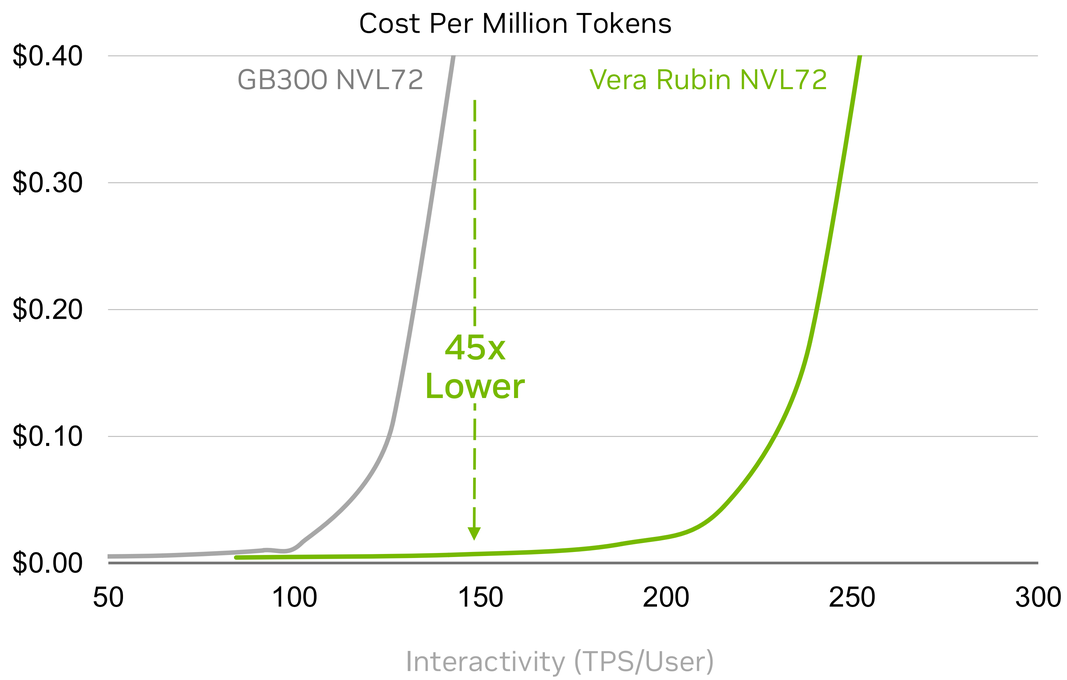
Inference Optimizations

- Disaggregated Serving
- KV Cach Offloading
- KV-aware Routing
- Topology Aware Scaling
- Kernel Fusion
- Low Precision Kernels
- Speculative Decode
- Compute-Comms. Overlap
- Fast Model Loading

Extreme Co-Design Delivers Performance Gains

Disaggregated Serving	Rate Matching	Large-scale Expert Parallel	Distributed KV-caching	KV-cache Offloading	KV-aware Routing	Fused CUDA Kernels





Optimize Model And Infrastructure For Your Usecase

Mix of open and proprietary models

Performant infrastructure right-sized to your use case

Lower tokens per task

Lower cost per token

Lower cost per task

Thank You!



OCP GLOBAL SUMMIT | 2026