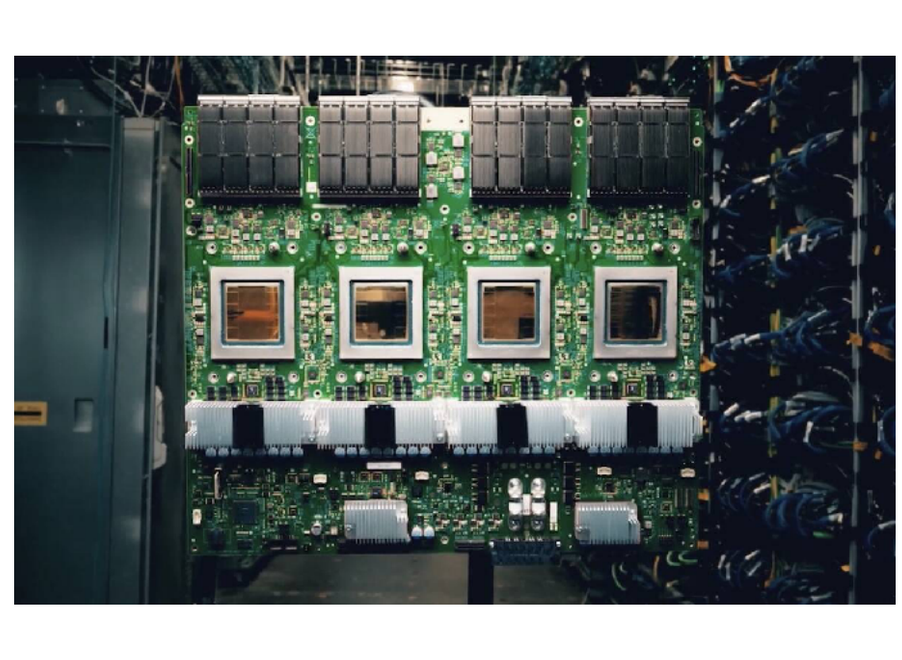
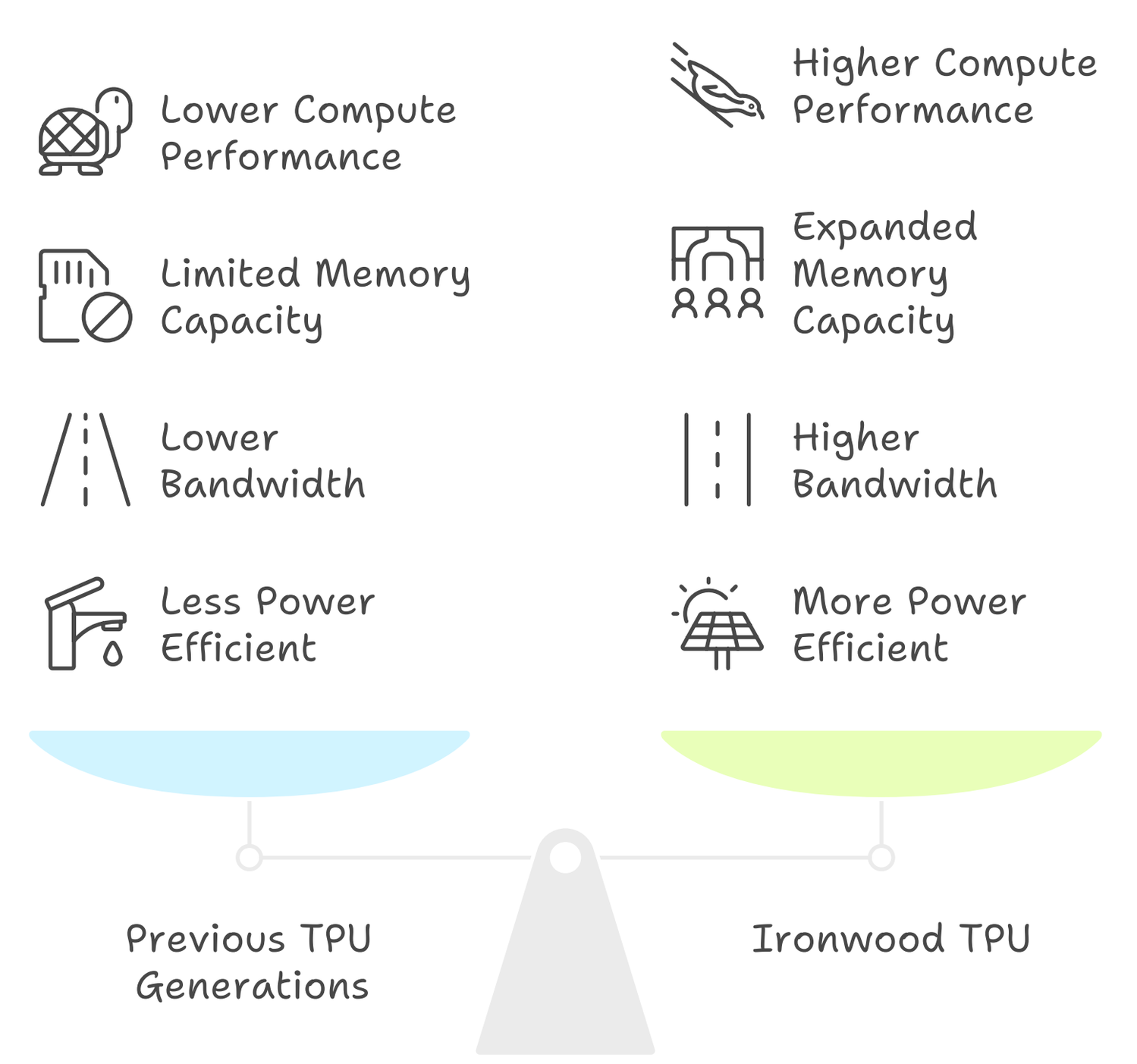


Rohit Singh Chouhan

Google Cloud Platform

Ironwood TPU (7th Gen) — "Age of Inference" Accelerator

Ironwood TPU enhances AI performance significantly.



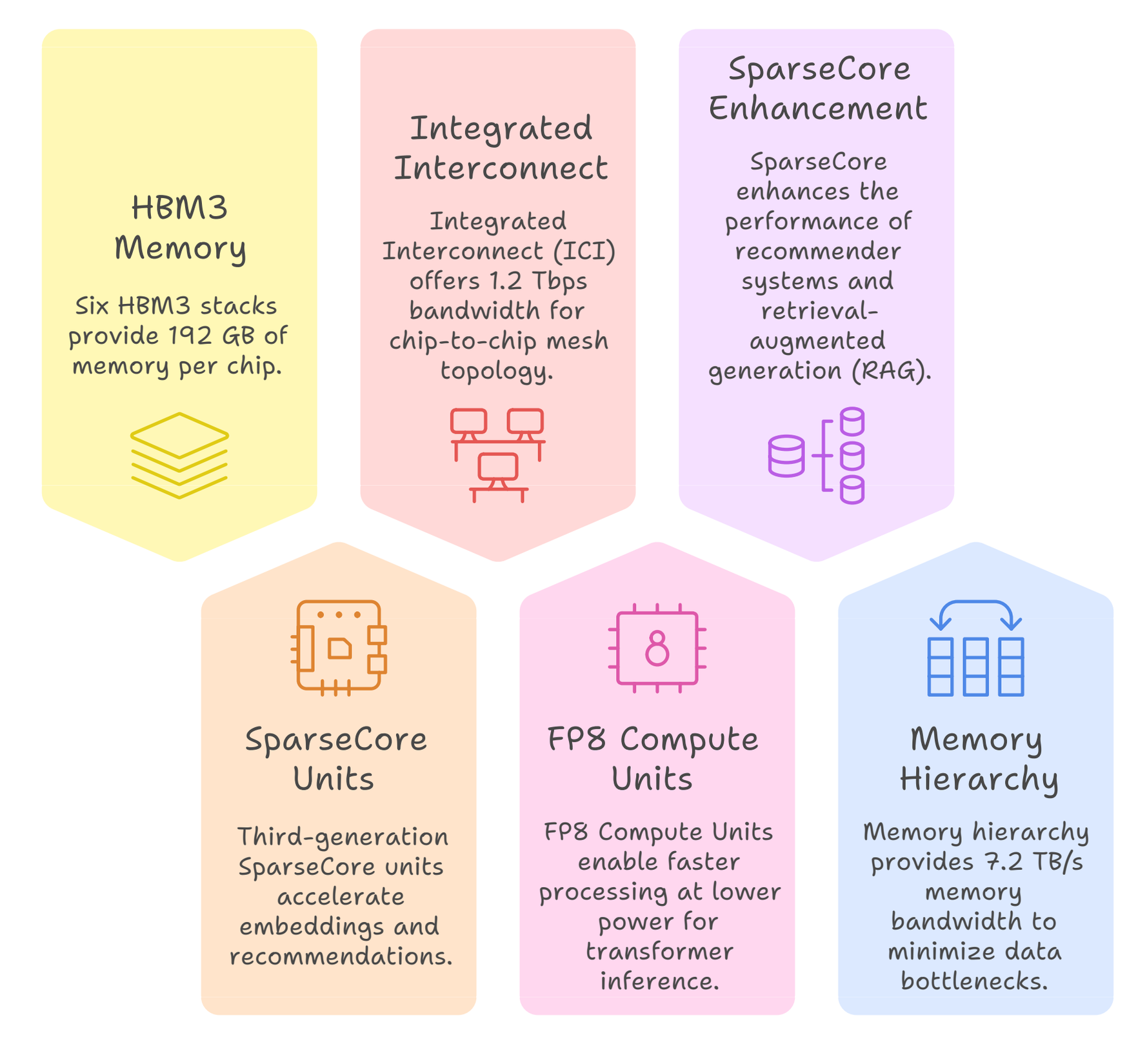
Evolution of TPUs

TPU Evolution

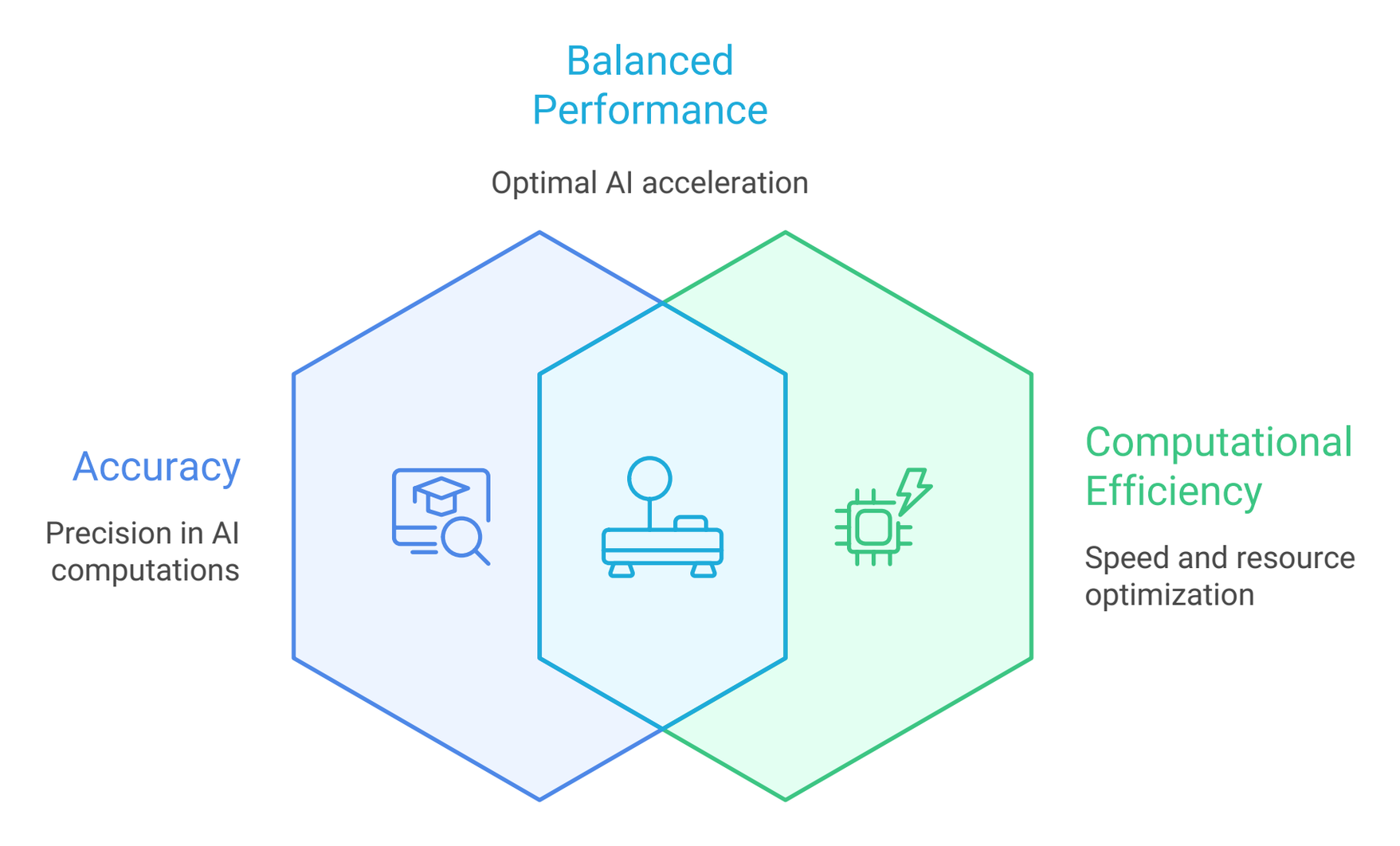
Generation	Primary Use Case	Precision	Key Features
TPUv1-v2	Early Inference	8-bit/16-bit	Efficient matrix computation
TPUv3-v4	Training and Inference	Higher	Greater memory bandwidth
TPUv5/v5p	Training LLMs	Mixed (bfloat16, int8)	Optimized for large language models
TPUv6e (Trillium)	Large Model Training	Unknown	Cost-efficient training
TPUv7 (Ironwood)	Inference-First	Unknown	Active reasoning and generation

Ironwood Architecture

Chip Features

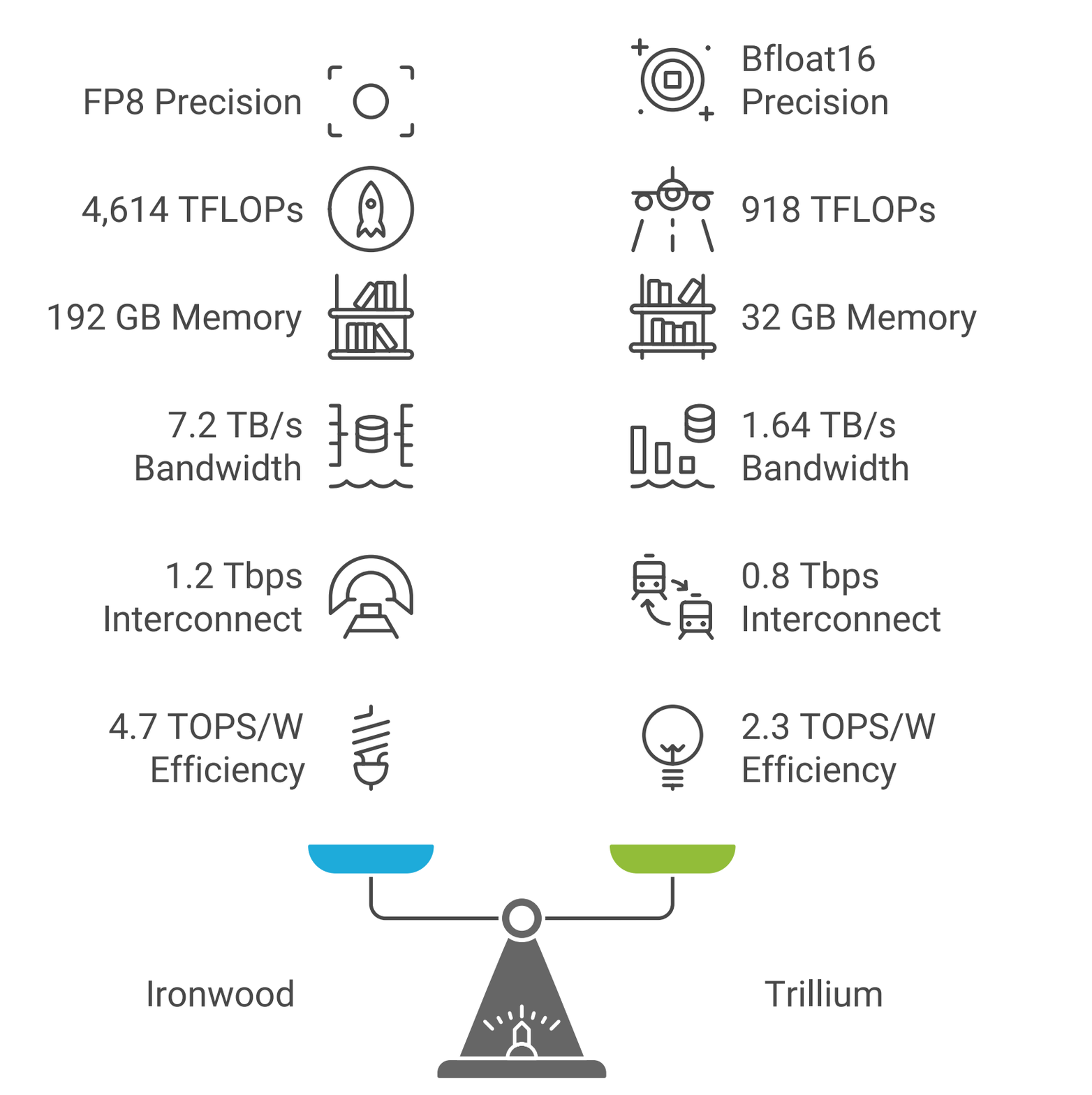


The Sweet Spot of AI Acceleration



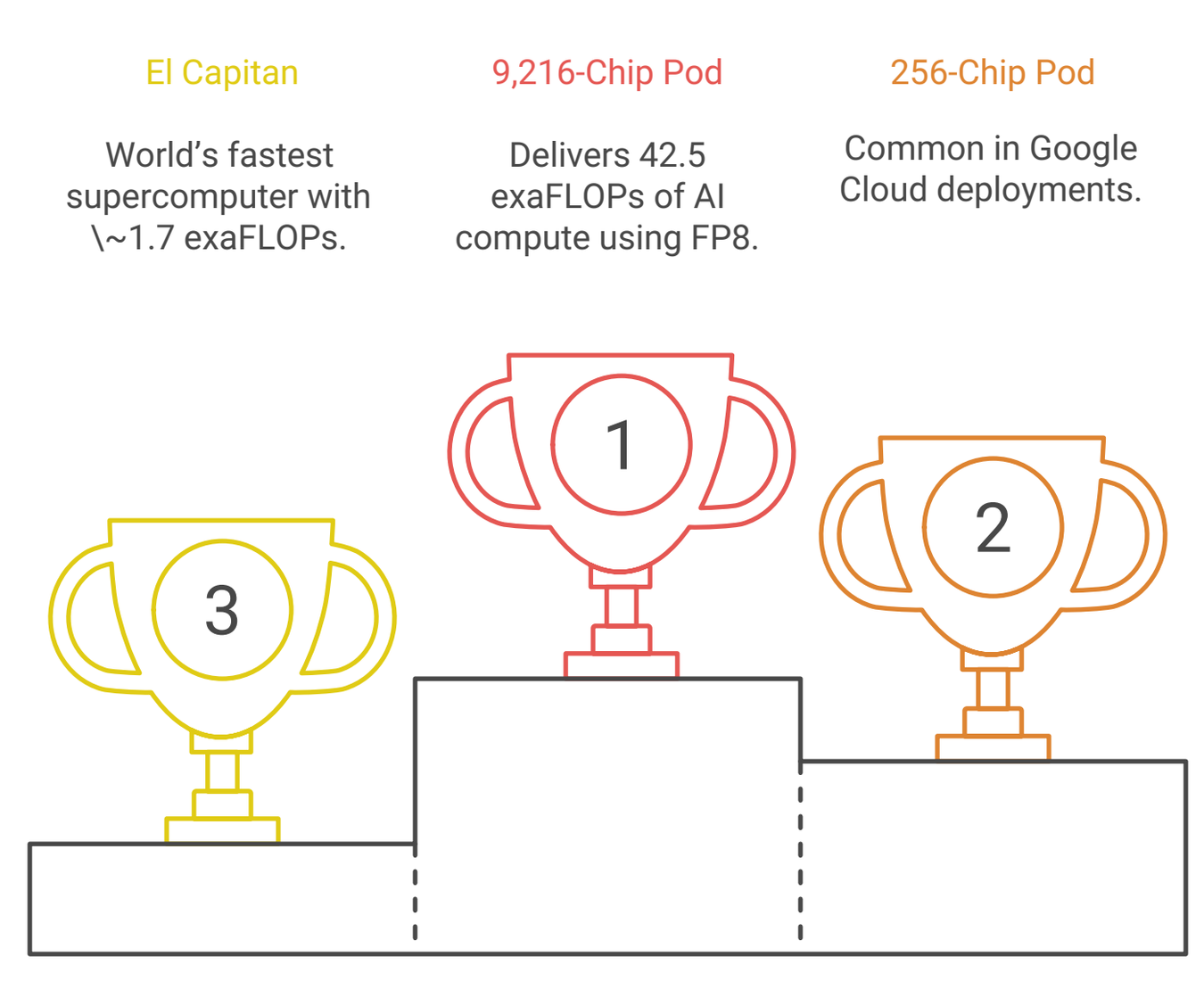
Hardware

Ironwood outperforms Trillium in key performance metrics.



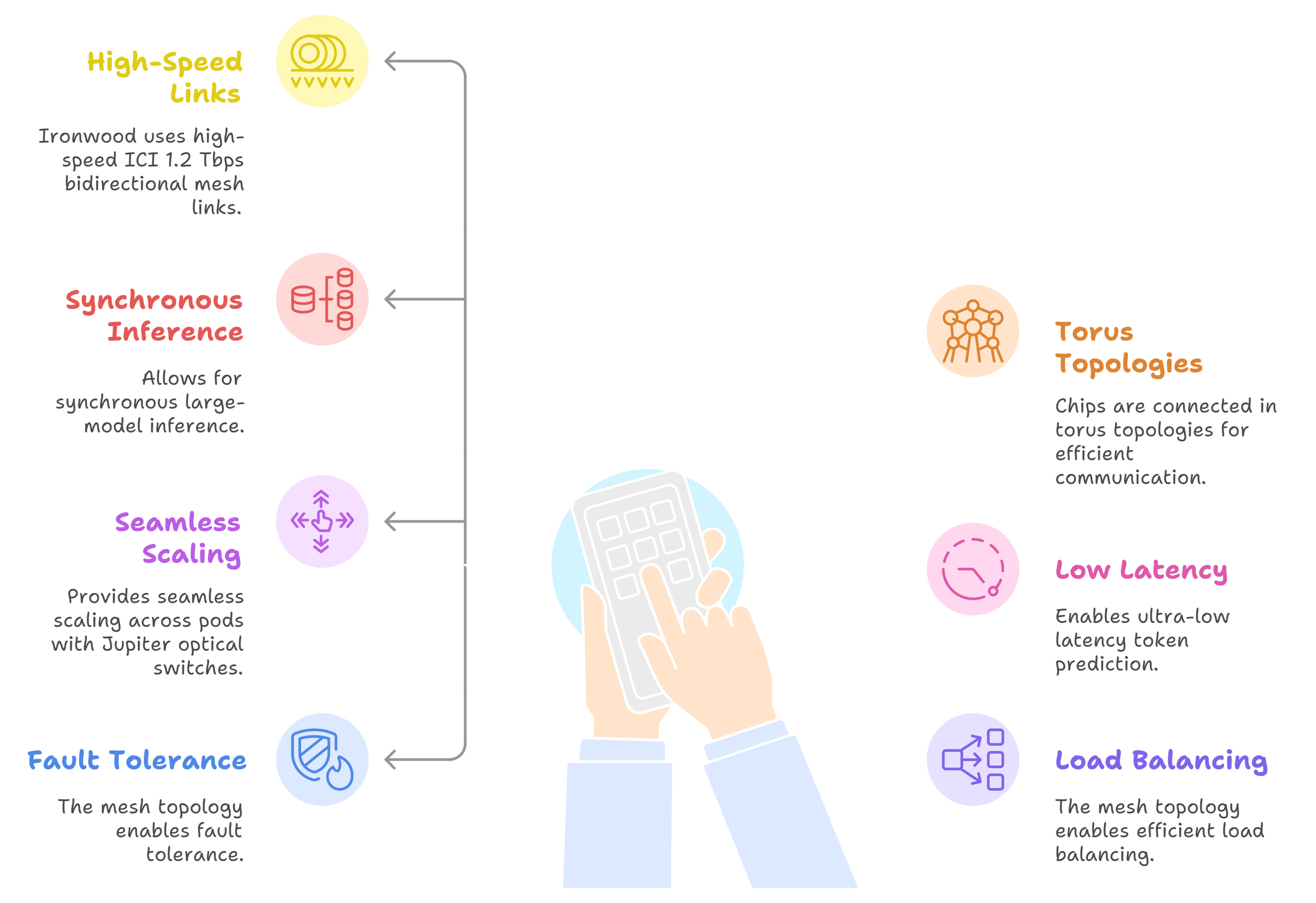
Pod-Level Design and Performance

Top AI Compute Pod Configurations



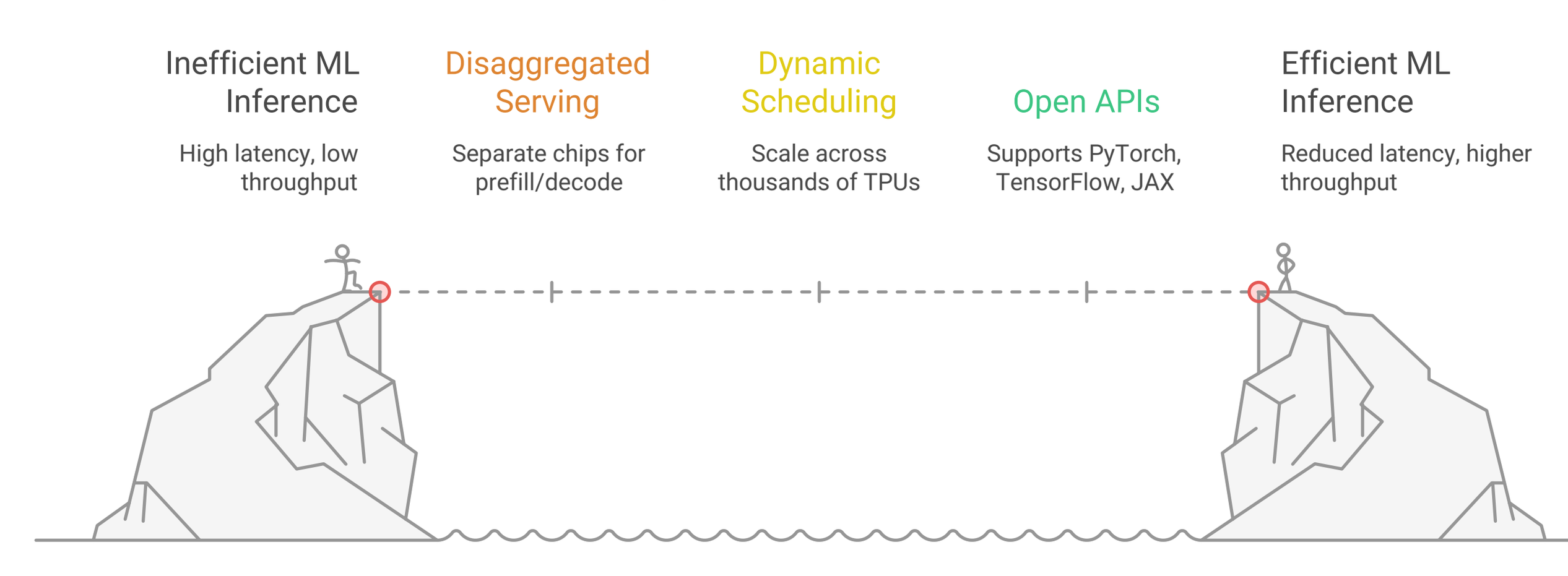
Interconnect & Topology

Ironwood Features



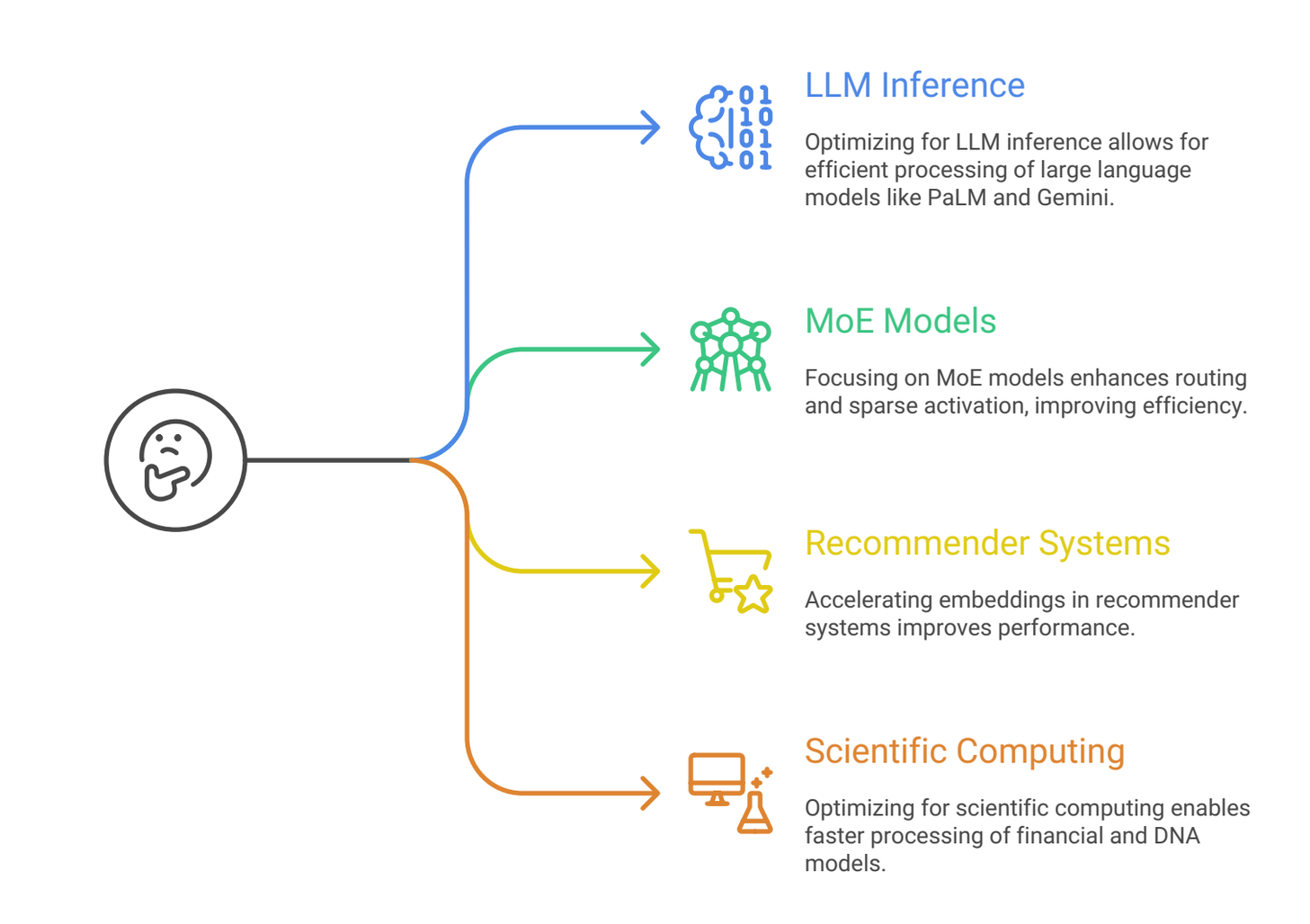
Pathways Runtime & Software Ecosystem

Pathways: ML Runtime Evolution



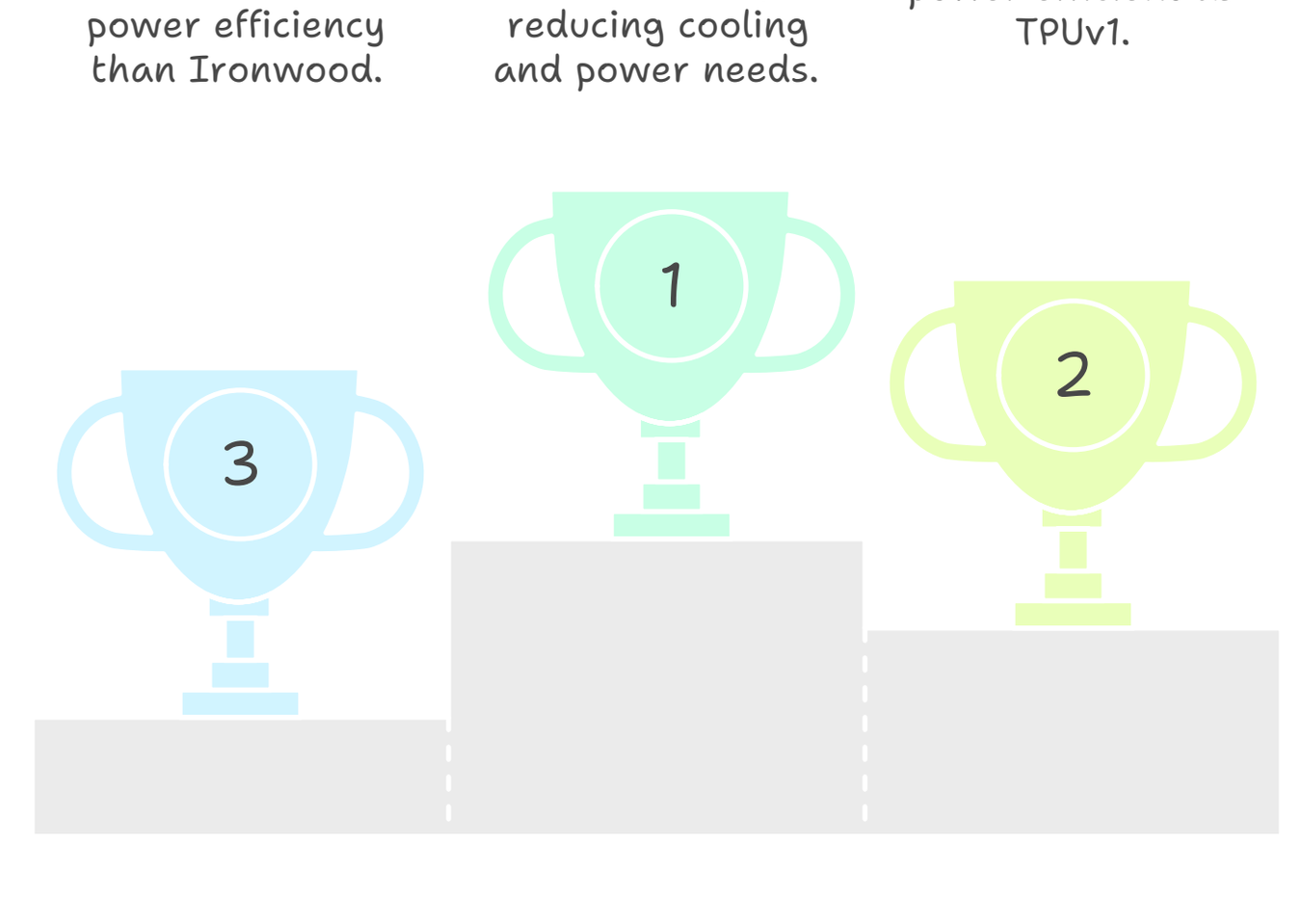
AI Use Cases Enabled

Which workload should Ironwood be optimized for?



Efficiency & Environmental Impact

TPU Power Efficiency Evolution



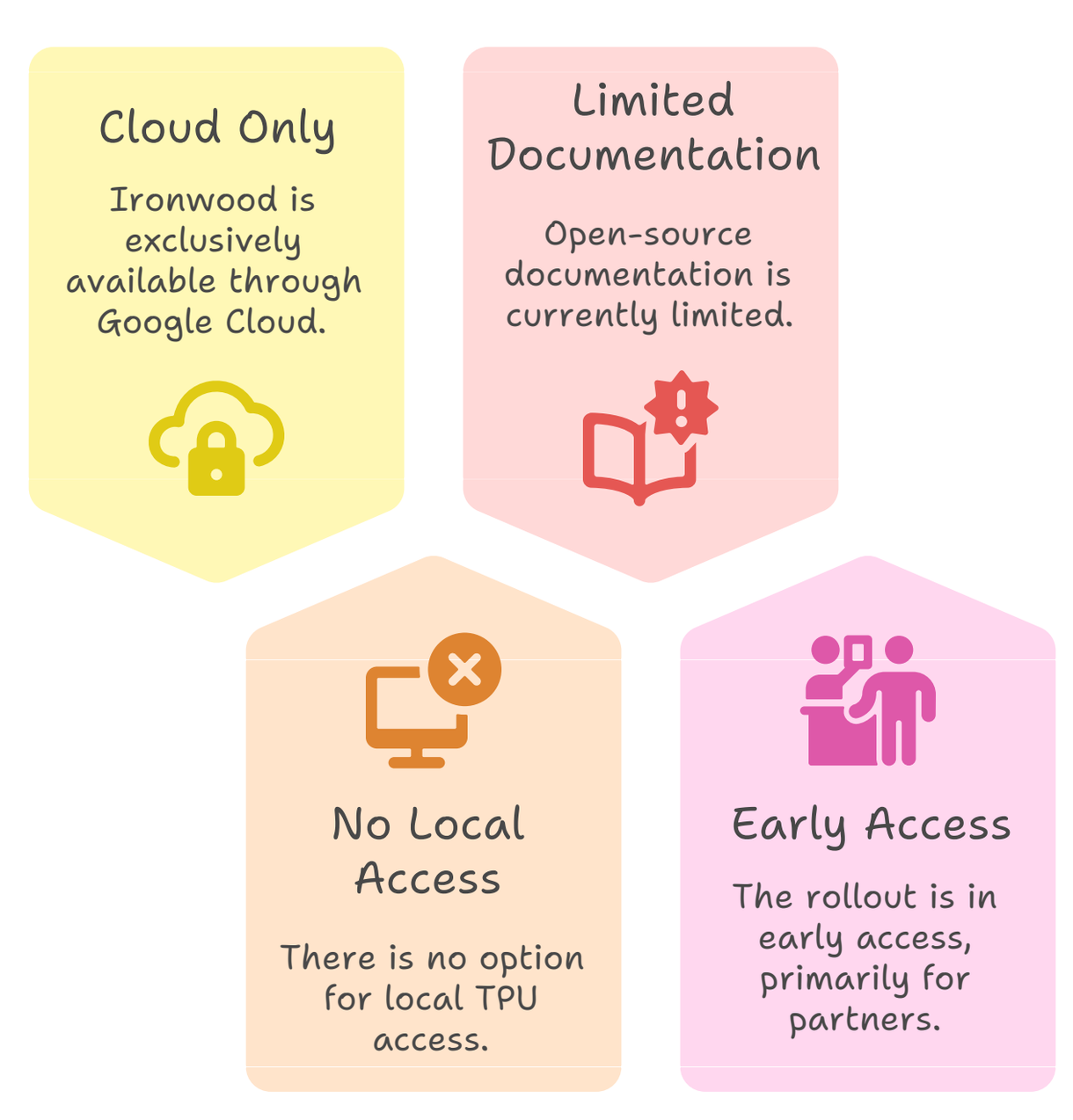
Competitive Comparison

Chip Comparison

Feature	Ironwood	NVIDIA H100	AMD MI300A
Peak FP8	4,614 TFLOPs	4,000 TFLOPs	1,300 TFLOPs
Memory	192 GB HBM	80 GB HBM	128 GB HBM
Interconnect	ICI 1.2 Tbps	NVLink 900 Gbps	Infinity Fabric
Efficiency	~4.7 TOPS/W	~2.5 TOPS/W	~2.1 TOPS/W

Limitations & Accessibility

Ironwood Limitations



Strategic Implications

Google's Vertical Integration Strategy