Ultra-fast AI Inference at the Edge

Frank Thiel
CTO of Gigantor Technologies
Silicon Catalyst Advisor
CEO of Kolvenier Solutions

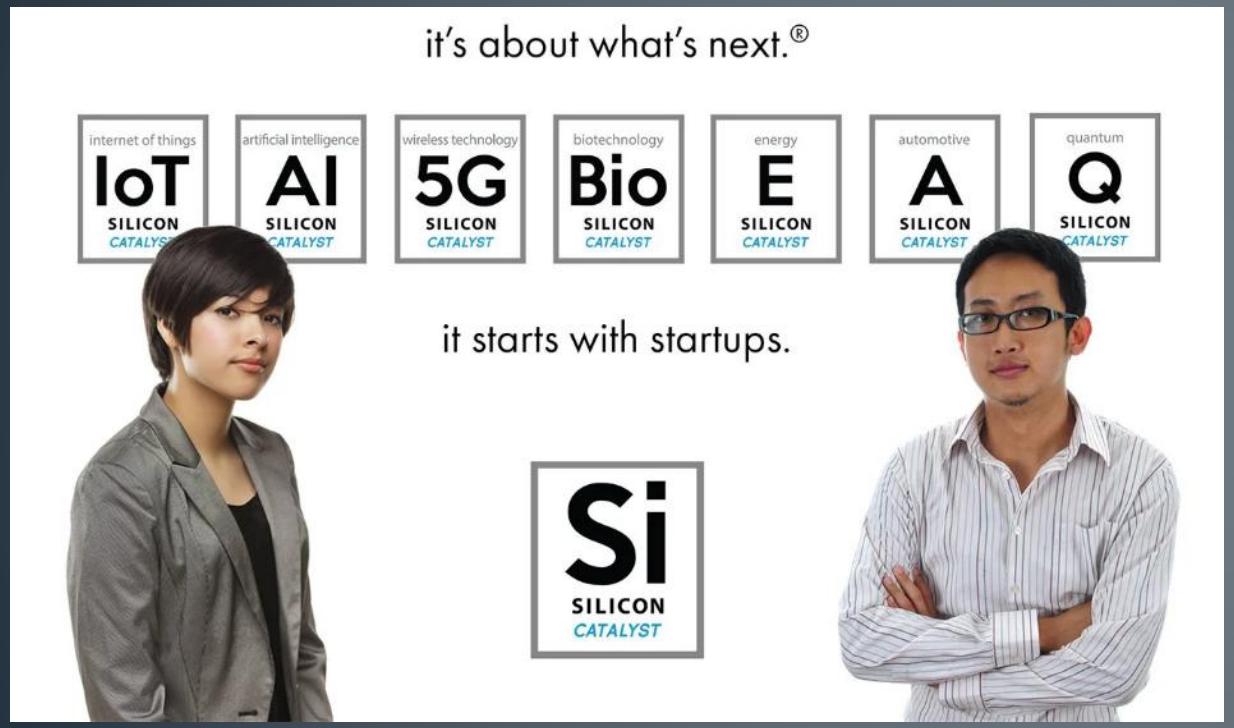
About Me

- Advisor for Silicon Catalyst
- CTO of Gigantor Technologies
  - High speed AI edge vision company
- CEO of Kolvenier Solutions
  - Specializing in guiding startups
- Over 4 decades in the semiconductor industry
  - Executive both as a GM and VP of R&D
  - Microsemi, Zarlink, Legerity, AMD, TI, IBM
- Hold 18 patents and am a published author.
- Started my career as a design engineer at Texas Instruments

GIGANTOR™ TECHNOLOGIES

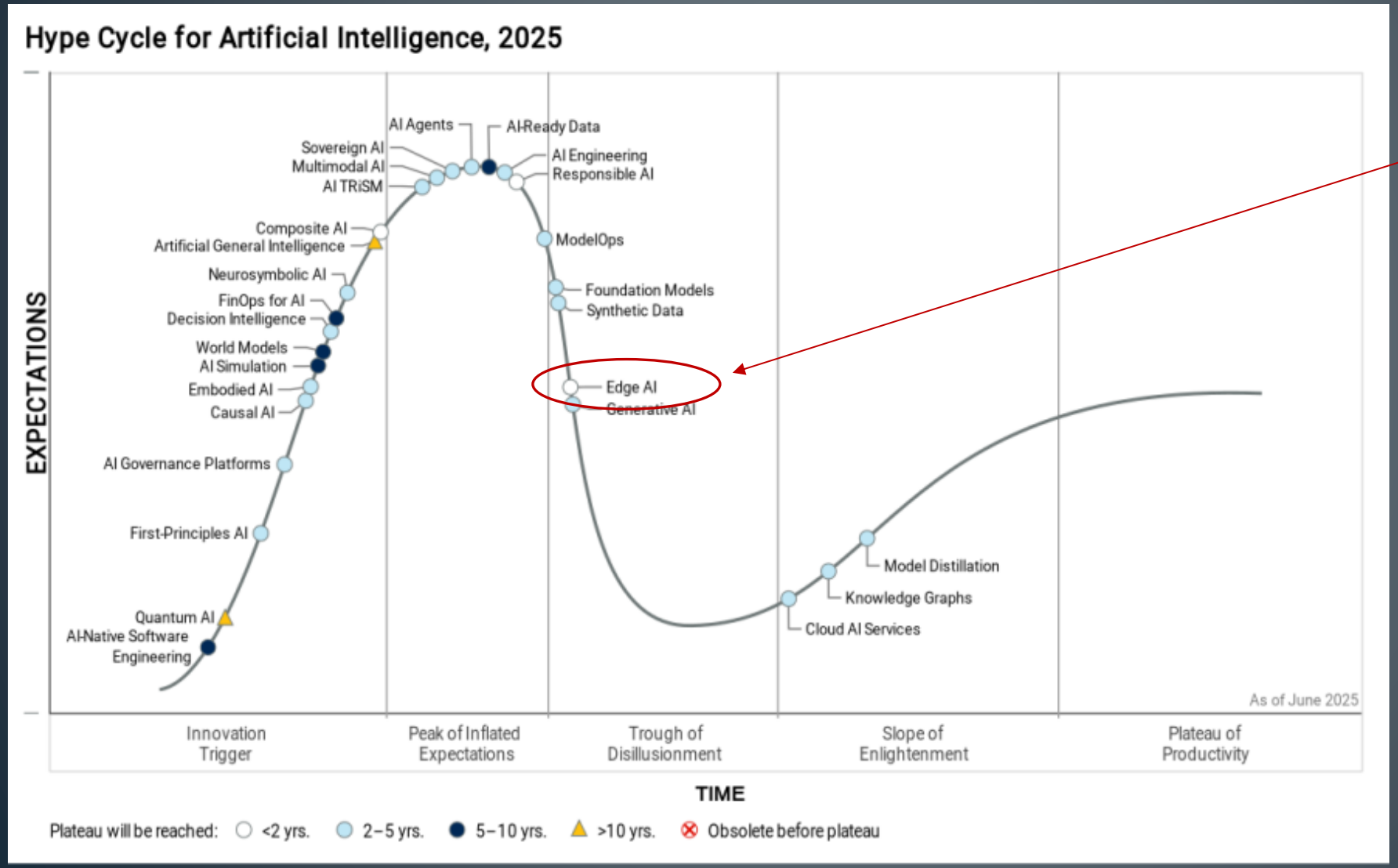
Kolvenier Solutions LLC
Business & Technology Consulting

Avisor → Member company



- Silicon Catalyst is the *only* incubator solely focused on semiconductors
- Being an advisor gives broad exposure to many innovative technologies and some of the most creative minds in the industry
- Like me, several advisors of moved into management roles with member companies

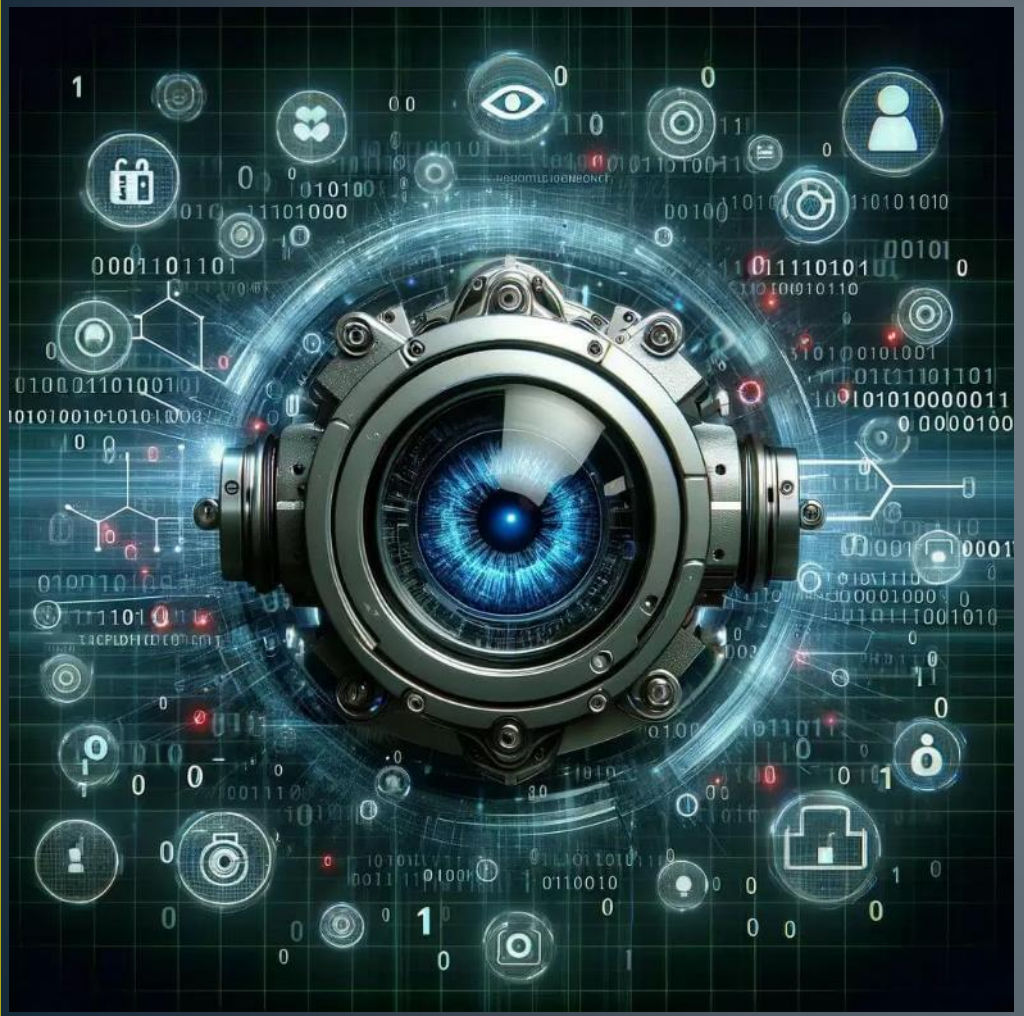
2025 Gartner Hype Cycle for AI



Edge AI reaching disillusionment – the precursor to productivity

Gartner

AI Training → AI Inference

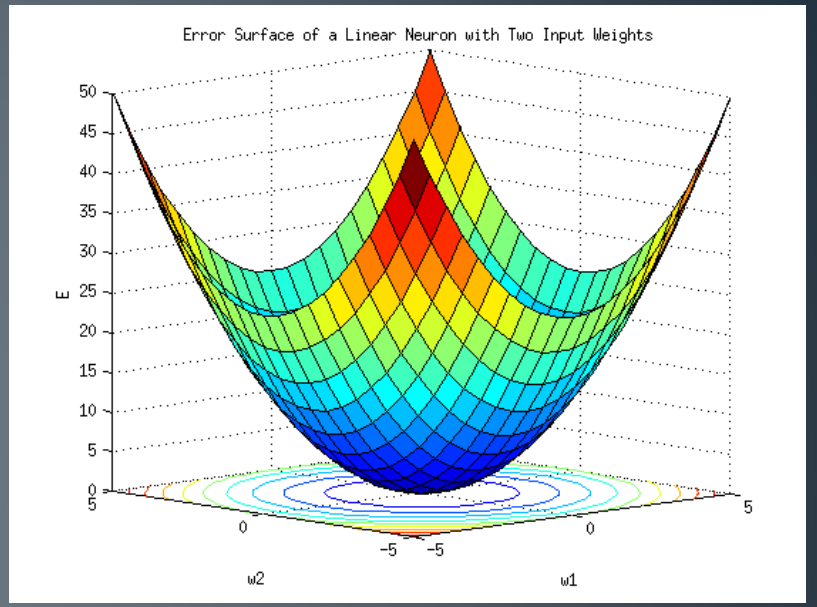


- Hyperscalers and AI players are spending billions of dollars on AI infrastructure/data centers
  - Amazon, Google, Microsoft and Meta spending more than $300 billion in 2025
  - OpenAI and Softbank have announced plans for investments of $500 billion over the next four years
- Training demand will be front-loaded into the next 2-5 years
- 80% of the demand now is for large-scale training facilities with only 20% on inference
- By 2030/32 we expect to see the ratio reversed
- **Training is a *cost – the ROI is from inference***

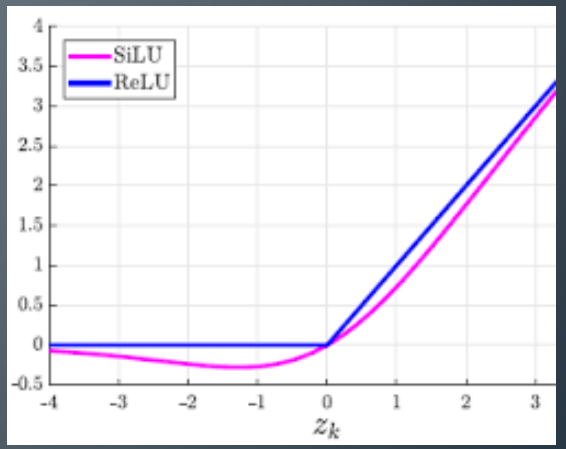
https://www.alvarezandmarsal.com/insights/rethinking-ai-demand-part-1-ai-data-centers-are-experiencing-surge-training-demand-what#
https://epoch.ai/blog/can-ai-scaling-continue-through-2030
https://fvivas.com/en/basic-guide-to-computer-vision-ai/#google_vignette

Inference vs Training

- Training typically requires full precision weights, biases, and models along with nuanced activation functions
  - FP32 weights and biases
  - Sigmoid Linear Unit (SiLU or Swish), Gaussian Error Linear Unit (GELU), etc.
  - Batch normalization
  - Iterative gradient descent and backpropagation runs
  - Power intensive!



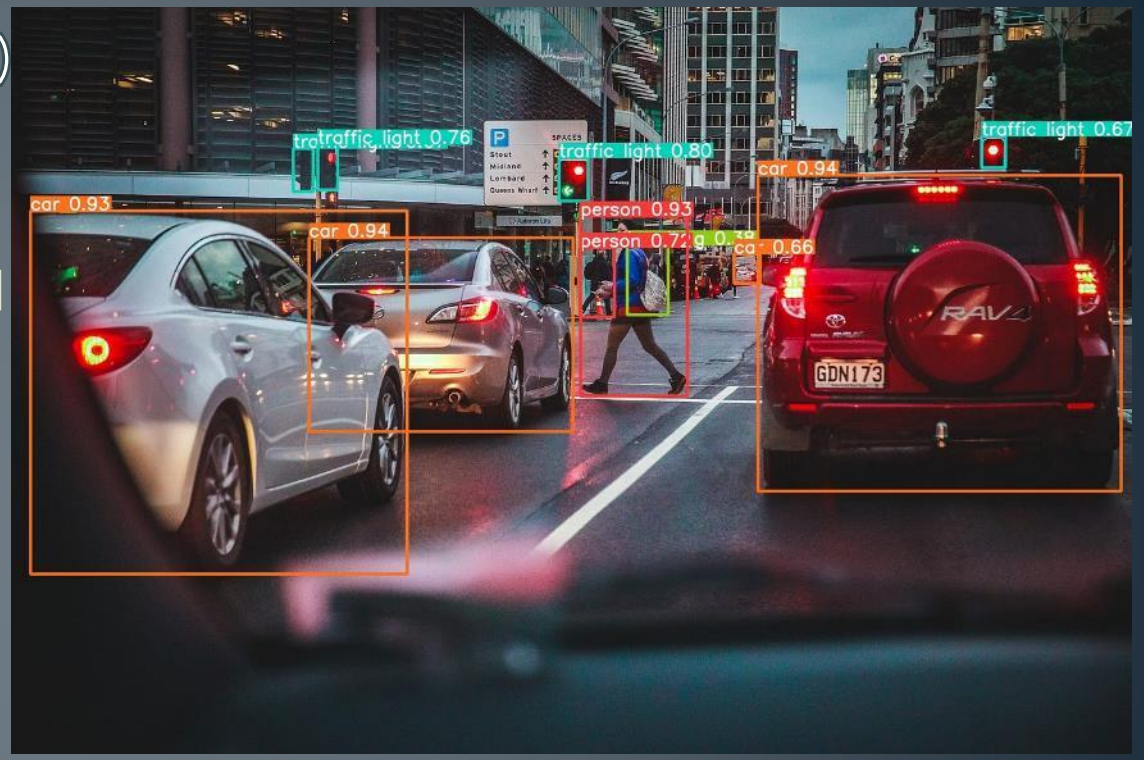
- Inference can use quantification and simplification
  - Reduced precision (INT8, or even INT4 weights and biases)
  - Bind normalization into the weights and biases
  - Simpler (less linear) ReLu activation



https://upload.wikimedia.org/wikipedia/commons/6/6d/Error_surface_of_a_linear_neuron_with_two_input_weights.png

Vision Inference Example: YOLO (You Only Look Once)

- Most popular multi-object tracking model today
- Anchored bounding boxes with “objectness” and class probabilities



- Faster inference speed and multiple scales/depths-of-field allow more objects to be tracked
  - More frames per second means better movement tracking
- Low power operation allows deployment in mobile devices

Convolutional Neural Networks (CNN)

- YOLO, ResNet and other popular vision models rely on CNNs
  - Different than Large Language Models like ChatGPT and Gemini
- Fundamental equation

$$o_{r,c,m} = b_m + \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{n=1}^{N} w_{i,j,n,m} \times a_{r+i-2,c+j-2,n}$$

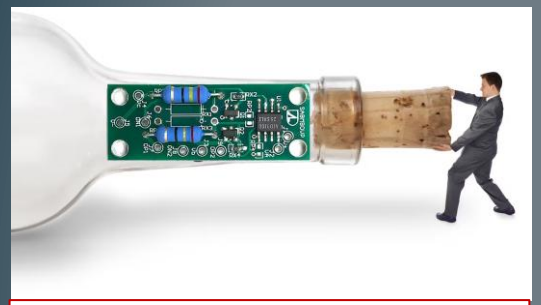
- Requires billions of Multiply-Accumulate (MAC) operations and fast, tightly-coupled memory (HBM, SRAM, etc.)

YOLOv8 Model	GFLOPs (Billion FLOPs) per 640x640 frame	Estimated MAC Operations (Billion MACs)	Model Parameters (Millions)
YOLOv8n (Nano)	8.7	≈4.35	3.2
YOLOv8s (Small)	28.6	≈14.3	11.2
YOLOv8m (Medium)	78.9	≈39.45	25.9
YOLOv8l (Large)	165.2	≈82.6	43.7
YOLOv8x (Extra Large)	257.8	≈128.9	68.2

Some edge inference hardware approaches

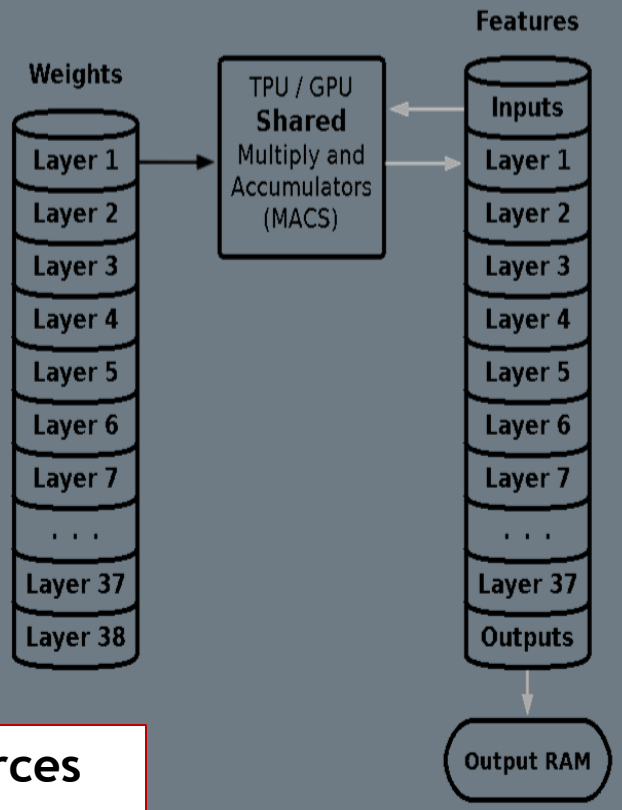
- Hailo -
  - Structure-Defined Dataflow with data pipeline for each task
- Groq
  - Introduced LPU in 2016, the first chip purpose-built for inference
- Mythic
  - Analog compute-in-memory purpose-built for AI inference
- NVIDIA
  - Single instruction, multiple threads (SIMT) added to GPUs
  - CUDA, Tensor cores
- EdgeCortix
  - Dynamic neural accelerator (DNA) engine, adding on-chip SRAM
- SiMa
  - MLA enhanced with hardware blocks to accelerate GenAI computations
- Gigantor
  - Fully deterministic hardware parallel pipeline for ultra-fast AI inference at the edge

Gigantor's Differentiators



Von Neumann bottleneck

GPU/TPU challenges

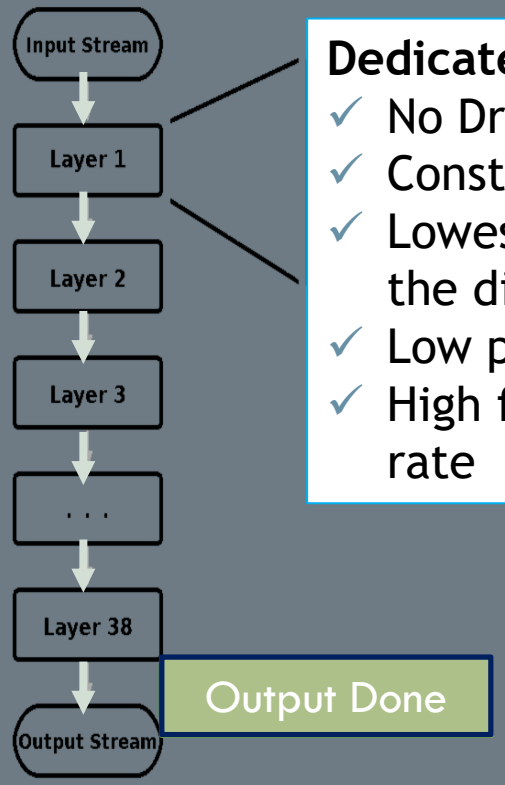


Shared hardware resources

- × Inconsistent latency
- × May drop frames
- × Higher power demands
- × Speed limited

https://www.electronicdesign.com/technologies/embedded/article/21156009/gsi-technology-breaking-the-von-neumann-bottleneck-a-key-to-powering-next-gen-ai-apps

Fully optimized parallel pipeline



Dedicated hardware

- ✓ No Dropped Frames
- ✓ Constant Latency
- ✓ Lowest latency possible in the digital domain
- ✓ Low power
- ✓ High frame resolution and rate

GIGANTOR™ TECHNOLOGIES

Thank You

Frank Thiel
CTO of Gigantor Technologies
Silicon Catalyst Advisor
CEO of Kolvenier Solutions

Email me at frank@kolvenier.com