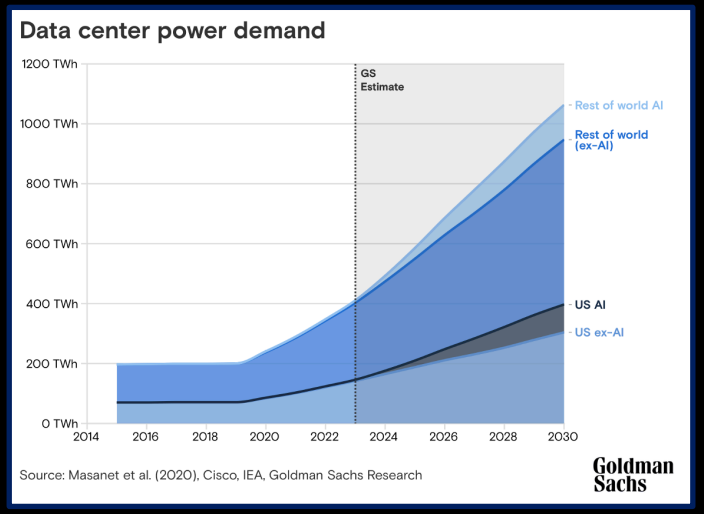
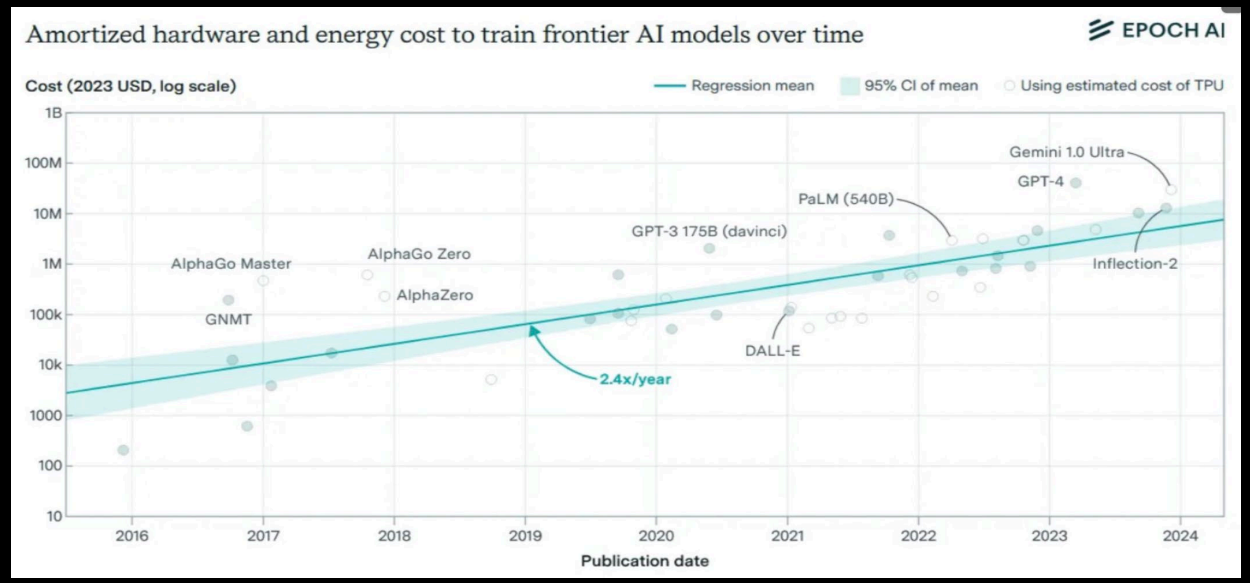
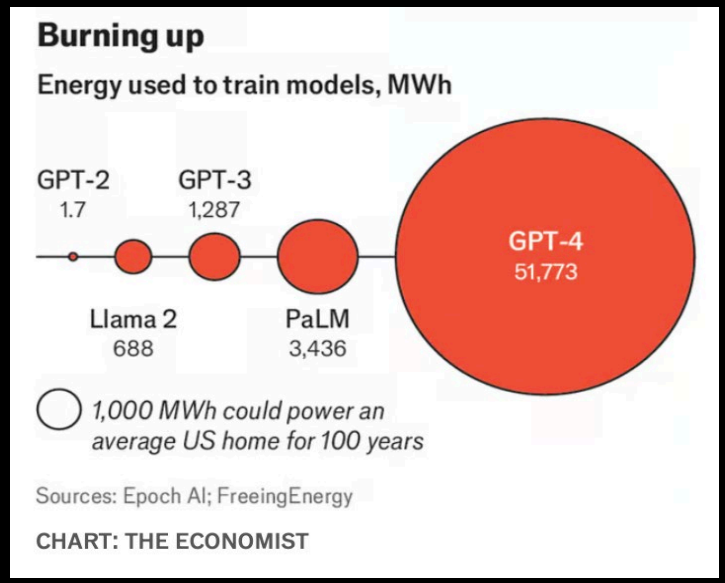
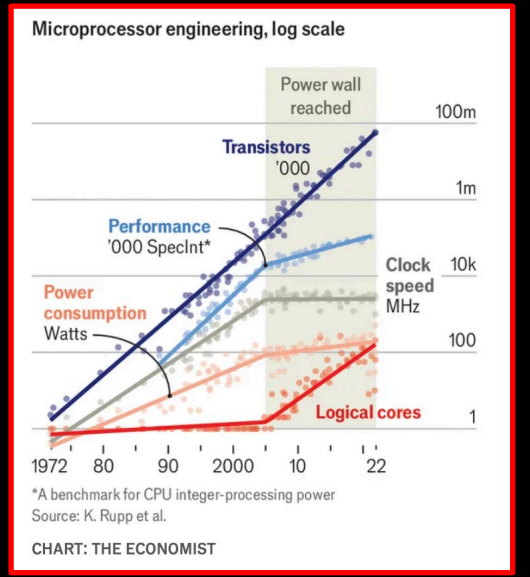
Increasing computational demands & the end of Dennard Scaling require innovation for performance & power efficiency





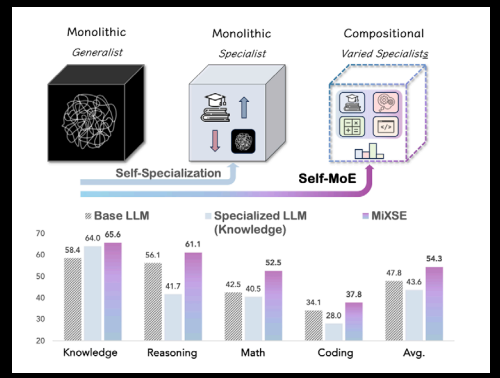




Full-Stack Innovation for AI energy efficiency

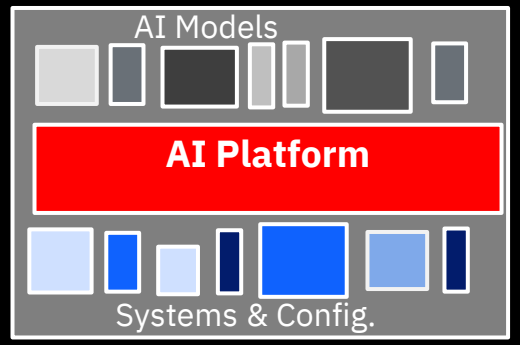
AI Models

- Bigger is not always better
- Reuse and collaboration for overall efficiency
- Architecture innovation: Mixture of Experts
- Beyond Transformer: Sequence or state space
- How we train and fine-tune matters



AI Platform

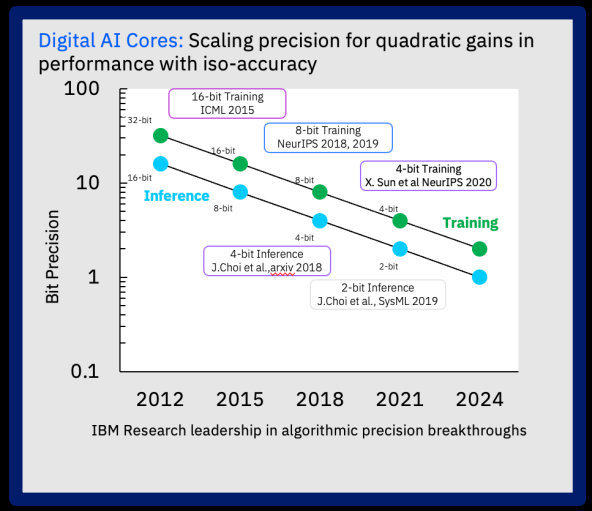
- Extensibility and Heterogeneity by design
- Built in Observability
- Transparency, Trust, and Community
- Optimization & scaling: run more with less



AI System Innovation

Spyre | **North Pole** | **Analog**

- Break the Von-Neumann bottleneck with innovative architecture
- Specialized ASICs
- Approximate computing
- Materials and Analog Research



Automotive, Embedded, Industrial, IoT

Enterprise, Data Centers

	0 Sensor Endpoint	1 Device Endpoint	2 Network/SFF Server	3 Edge Server	4 Enterprise Server/On-Prem Cloud	5 Public Cloud
Power	< 1W	<= 5W	<= 75W	<=75W	75W-350W	350W+
TOPs[2]	<1	5 - 20	20 – 50+	50+ – 300+	300 - 500+	300 - 1000+
Precision	INT2, INT4, INT8	INT4, INT8, Mixed, BF16	INT8, Mixed, BF16, FP16	INT8, Mixed, BF16, FP16	INT8, BF16, FP16, FP32	INT8, BF16, FP16, FP32, FP64
DRAM	LPDDR4	LPDDR4	LPDDR4	LPDDR4, 5 (x)	LPDDR4, 5 (x), GDDR6, HBM	GDDR6+, HBM3 (e)
FF GTM	IP, ASIC, SOC, SIP, MCM	ASIC, SOC, SOM, M.2, SFF PCIe	SFF PCIe, EDSFF	SFF PCIe, EDSFF	PCIe	PCIe, SXM, Custom

AIU NorthPole

AIU Spyre

GPU Control	Spyre Measured
0.6 img/sec/W	2.1 img/sec/W

70 TB Incoming Satellite Data Stream Daily

50+ Accelerators Continuously Required @FP16

System-Level Impact *Cards, Node Overheads, Cooling*

23 kW Power Saving (~20 US Homes) with AIU

85 Tons CO2 Equivalent Yearly Footprint Reduction

2x Further Savings Projected for next iteration