Scaling to Meet The Needs of AI

Pradeep Dubey

Intel Senior Fellow and Director, Parallel Computing Lab

EDPS Symposium

October 4th, 2024

intel®

Two Virtuous Cycles Accelerating AI Innovations

Expanding Application Space of AI

Making AI Algorithms More Capable

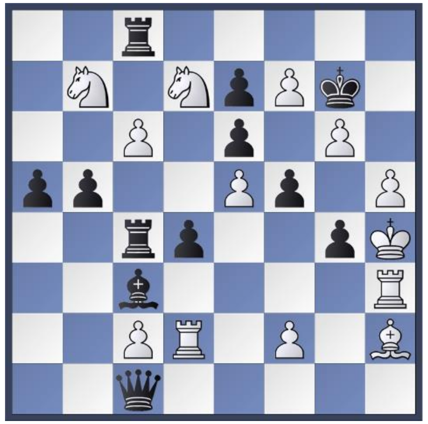
Designing Better Systems *for* AI

Designing Better Systems *with* AI

Ensuring AI Stays Good

Where are we headed next …

Making AI Algorithms More Capable: From *perception* to *reasoning*



AI Problem Statement & the GenAI Opportunity

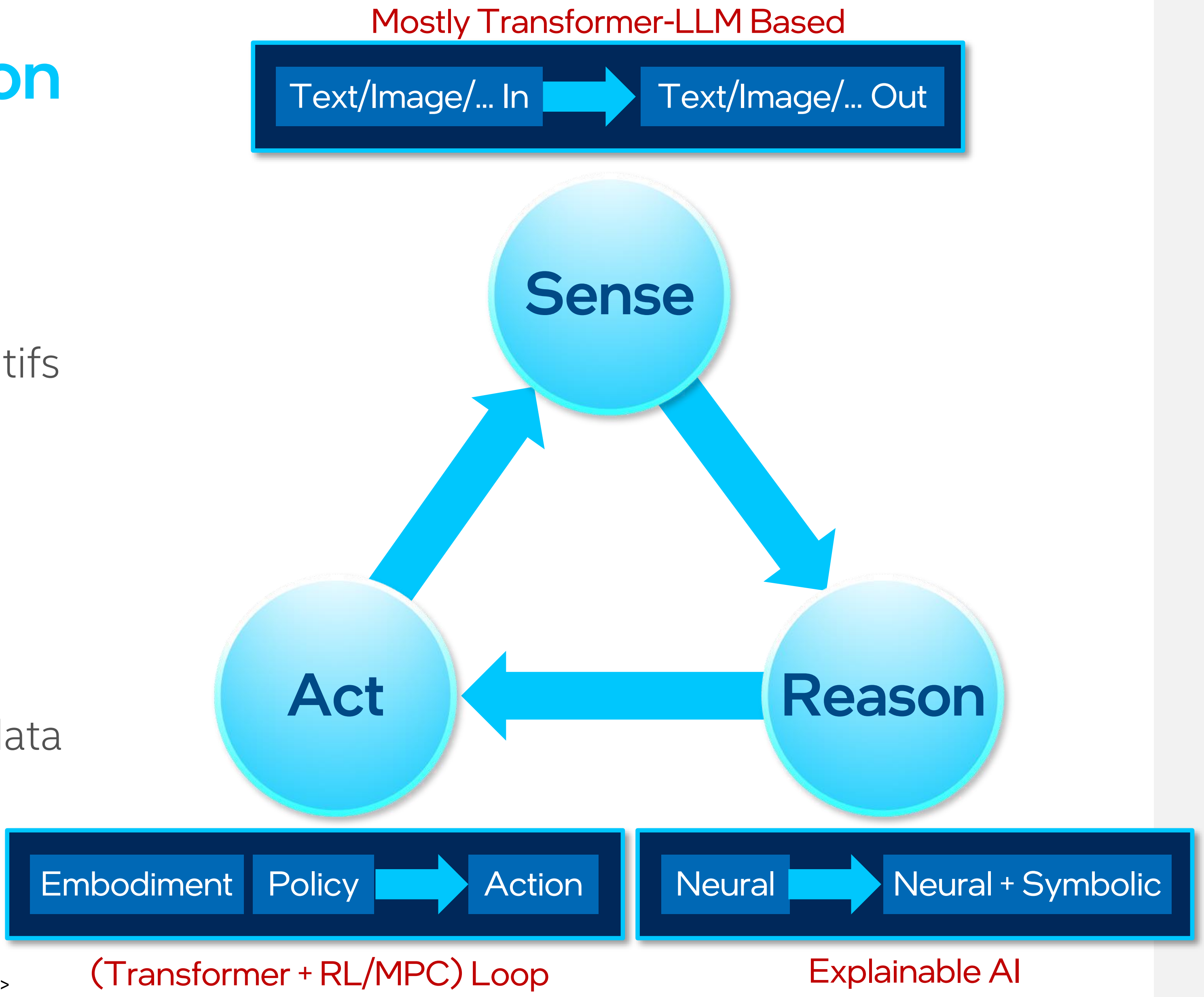
- AI Update:
  - Trained on text, image, code or protein … data, AI can now learn a model that generates a novel text, image, code or protein … that is indistinguishable, yet original compared to the training data

    → *AI's Dream Moment*: AI claiming to do what human minds do every night while dreaming
  - We can avoid training-from-scratch for various use cases, and instead cheaply (in cost terms) fine-tune or augment a large model down to a smaller-and-better model for a specific context.
- AI Opportunity:
  - Above two have significantly sped up the permeation of AI into various content creation and decision-making workflows, which were reluctant in adopting AI so far, simply because of economies-of-scale associated with *more machine-and-less-human-cycles* in the workflow.

Agentic Era AI

From Perception To Action

- Promises:
  - Kahneman's (S2 → S1) loop
  - Learned object features → action motifs
  - Multi-agent collaboration
  - Human-in-the-loop → Humanoids*
- Challenges:
  - LLM models running out of training data
  - Current algorithms need revisit **
  - Safe Superintelligence?

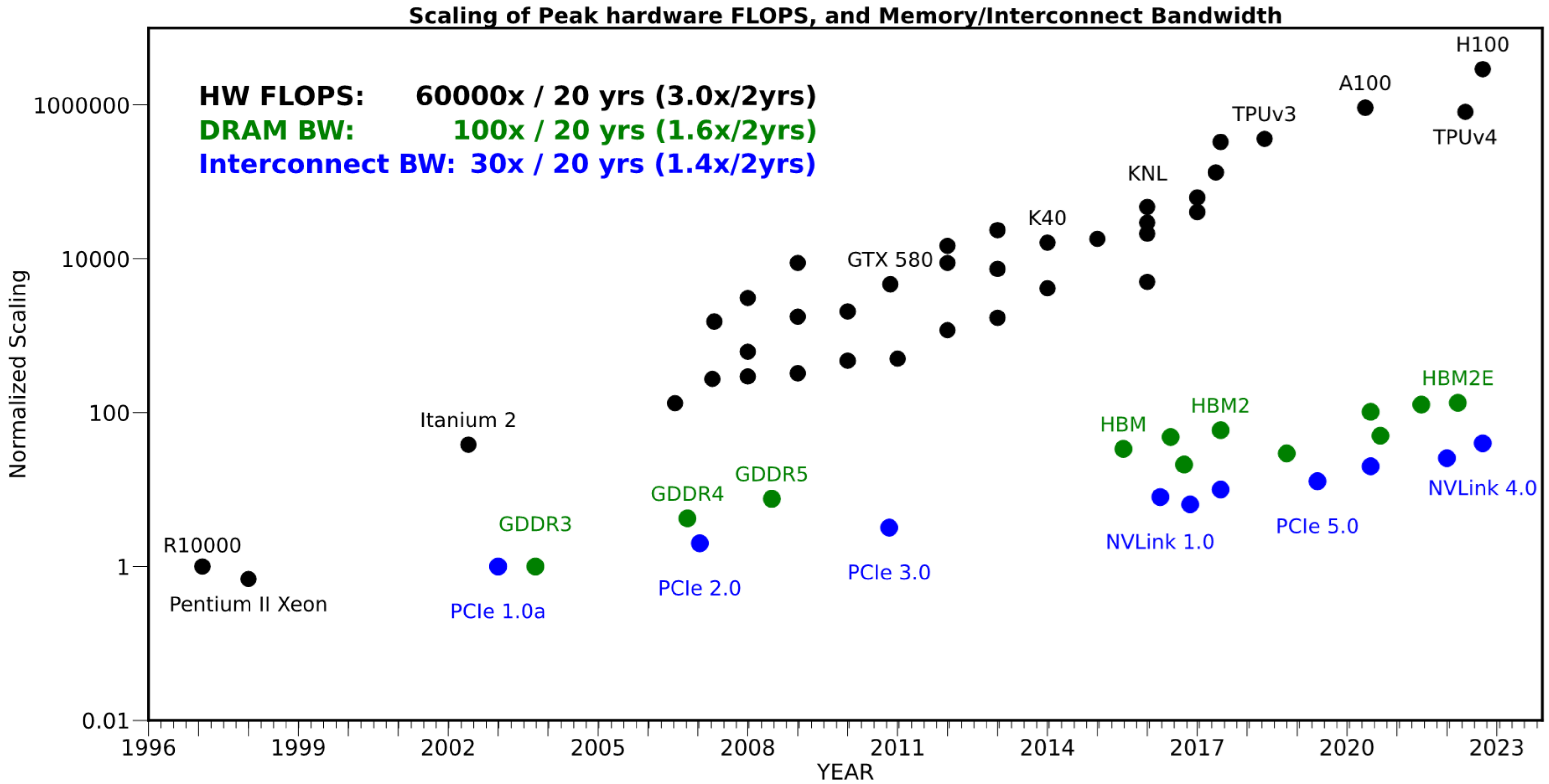


- Generally Capable Agents in Open-Ended Worlds, Jim Fan, NVIDIA GTC'24 <link>

** Objective-Driven AI, Prof. Yann Lecun <link>

Scaling Challenges In the Agentic AI Era

- Compute flops-bandwidth requirements have gone up 100-1000x or more at the high end (pre-training) and 10-100x for the volume server and client as well.
- Easier to meet compute flops needs than feed needs
- Highest-level manifestation of limitations that could end-it-all:
  - Energy cost of sustaining AI compute needs: primarily rooted in cost of data movement
  - Developer disconnect: Growing gap of ninja-performance vs. data-scientists

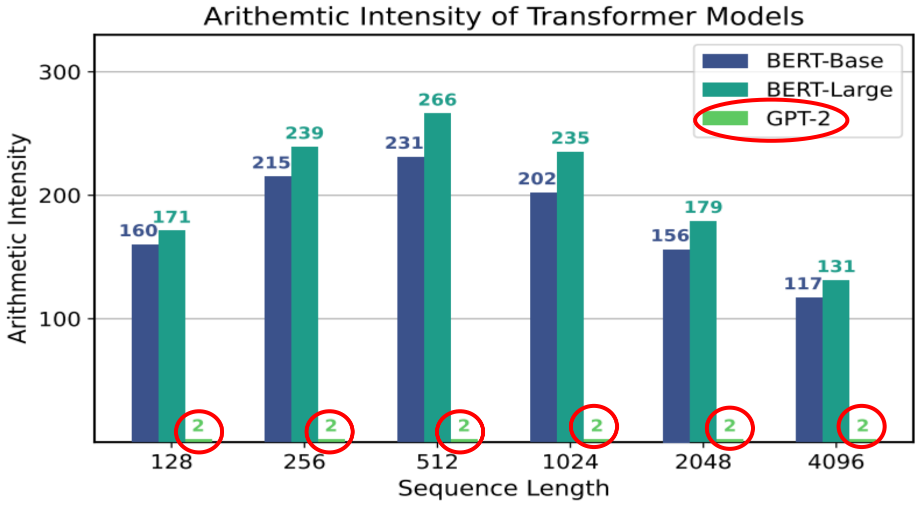


Credit: AI and Memory Wall, Amir Gholami <link>

Innovation Opportunities

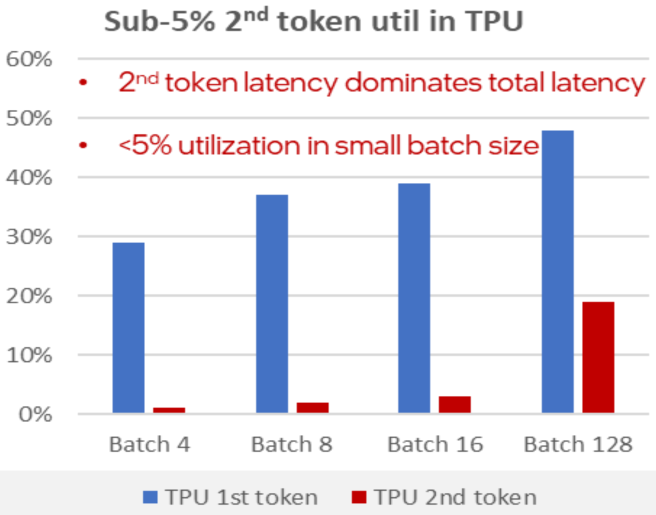
- Compute:
  - Numerics (~1b/tensor), unstructured sparsity, compute-in/near-memory/network, dataflow
- North-South: Feed the compute
  - >10x BW than HBM at some reasonable capacity-tradeoff at iso-power
- East-West:
  - Scaling up/out: High-radix, optical networking with <10fj/b-mm vs. >100fj/b-mm today
- Software:
  - Compiled performance, self-organizing code,
  - Natural language → SQL/plans … (AI generated)
- Packaging-and-cooling

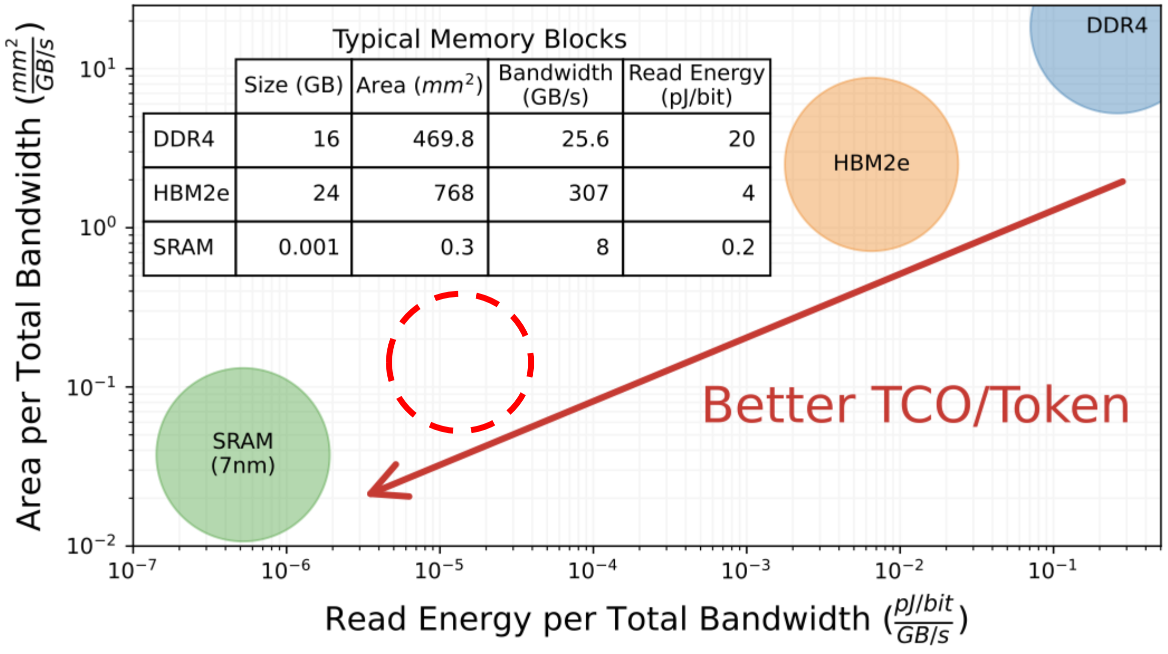
LLM Inferencing: Compute Intensity Challenged



Source: https://arxiv.org/pdf/2302.14017

Source:
Google [MLSYS ‘23](#)

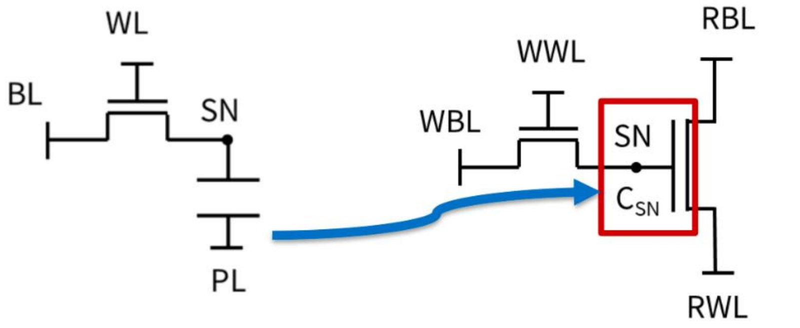




Source: Chiplet Cloud … <[link](#)>

Addressing Capacitor Limitation of DRAM *

1T1C DRAM **2T Gain Cell***



Storage capacitor → Gate capacitance of read transistor

*Gain cell: Shukuri, Kure and Nishida, IEDM 1992.
P. Meinerzhagen et al. Cham, Switzerland: Springer, 2018.

6 | H.-S. Philip Wong Shuhan Liu, *Stanford SystemX Alliance Ser*

TABLE I

GEMTOO Memory Macro Simulation Based on Simulated 28 nm ALD ITO FET Indicates That OS–OS GC and HGC Have Long Retention and High Frequency

28nm node, V_{DD} = 0.9V, sub-array size 64 row x 256 col.				
	SRAM[26]	Si GC# [7]	OS GC#	HGC#
Cell size* (μm²)	0.16	0.14	0.14/N	0.06
Refresh Period		19 μs	9 s	9 s
Max Freq. (MHz)	735	242	345	721
Bandwidth (GB/s)		7.6	11	23

Simulated with GEMTOO
* SRAM -- pushed design rule; GC -- logic design rules. For OS gain cell in this work, equivalent cell size depends on 3D stacking number (N) of layers.

* Design Guidelines for Oxide Semiconductor Gain Cell Memory on a Logic Platform, Phillip Wong, et.al. , IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 71, NO. 5, MAY 2024

Network Scalability

Era of '*Datacenter as a Computer*' is almost here

2013

The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second Edition
Luiz André Barroso, Jimmy Clidaras, Urs Hölzle
Morgan & Claypool Publishers (2013)

Datacenter today has supercomputer-class compute …

Image courtesy: Aurora Supercomputer, Argonne National Lab <link>

Animation Credit: What's next for photonics-powered data centers and AI; Nick Harris, CEO Lightmatter <link>

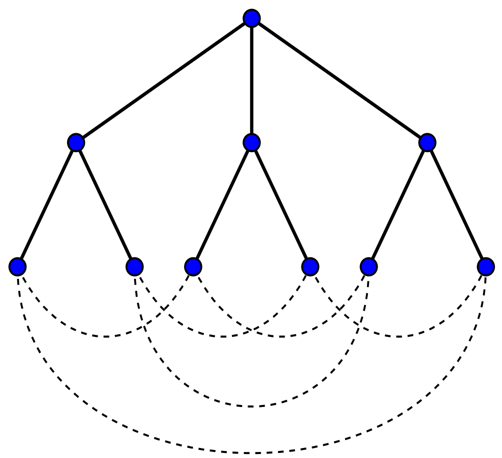
High Radix Networks and Photonics

- High radix, low diameter networks are technically a superior option …
  - The more routers a packet goes through, more the latency, congestion, delay …
  - Hop minimization should therefore be a goal: ideally 2 to 3 hops maximum

- Photonics offers a promising technology path for high-radix networks
  - Optical IO latency is determined by time-of-flight (ToF): 5 nsec/meter
  - Rack level ~10 nsec, data center with 40 meters: ~200 nsec
  - Total network latency = network hops * routing delay + ToF < 500 nsec

High Radix Networks and An Open Graph Theory Problem

- Degree/diameter problem:
  - What is the order of the largest graph with a given degree and diameter?
- Moore Bound limits the number of nodes in a graph:
  - d is the graph degree, or *router radix*
  - k is the diameter, or *maximum number of hops*.
  - $M(d,k) = \frac{\left((d)(d-1)^k\right)-2}{d-2}$
- Want to maximize nodes and minimize diameter
- It remains an open problem whether or how close one can get to the Moore bound in terms of a *constructible graph*

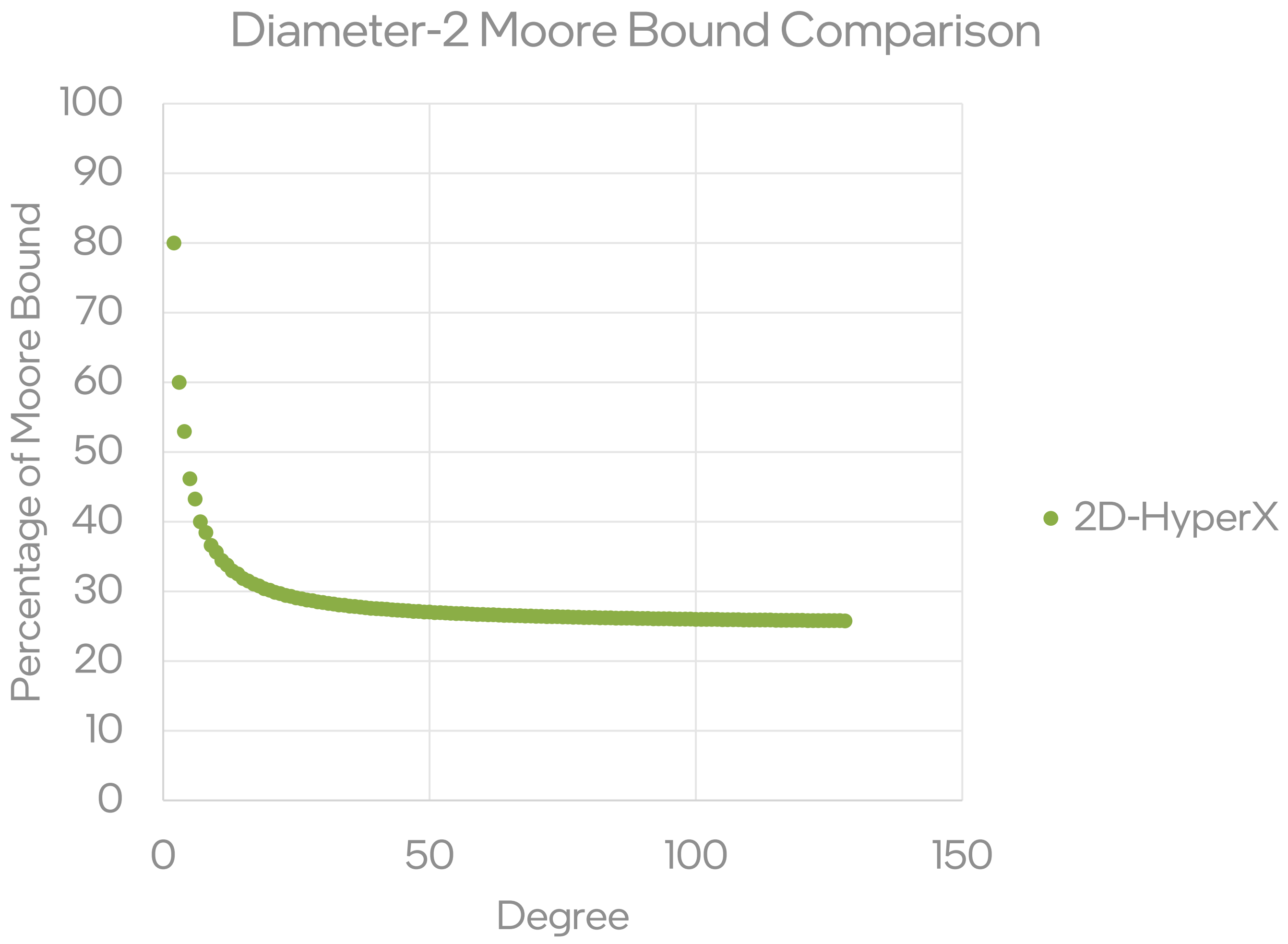


Moore Bound construction:
degree d = 3, diameter k = 2

Moore's Bound and the Degree/Diameter Problem

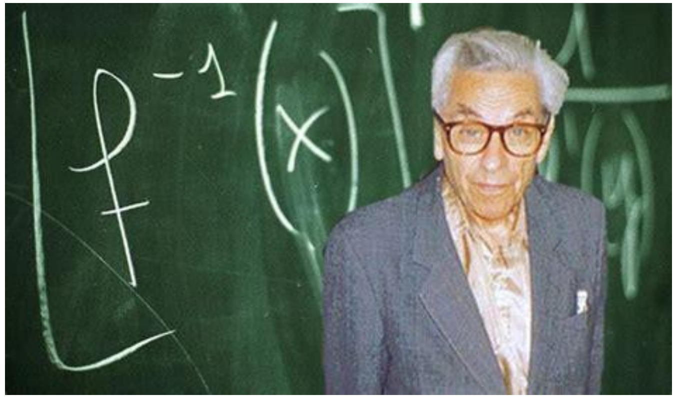
- The Moore bound is $\mathcal{O}(d^k)$.
  - 1 hop $(k = 1)$ can reach at most d different destinations.
  - 2 hops $(k = 2)$ can reach at most $(d^2 + 1)$ destinations.
    - For d=32, the bound is 1025.
    - For d = 64, the bound is 4097.
    - For d=128, the bound is 16,385
  - 3 hops $(k = 3)$ can reach at most $(d^3 - d^2 + d + 1)$ destinations.
    - For d=32, the bound is 31,777.
    - For d=64, the bound is 258,113.
    - For d=128, the bound is 2,080,897

Reality Today of Low-Diameter, High-Radix Networks: 2D HyperX



Erdős-Rényi polarity graphs …

- Discovered independently by Erdős-Rényi (1962) and by Brown (1966).
- An isomorphic construction was discovered even earlier by Singer (1938).
- ER graphs are from projective geometry over the finite field of order q.



Paul Erdős

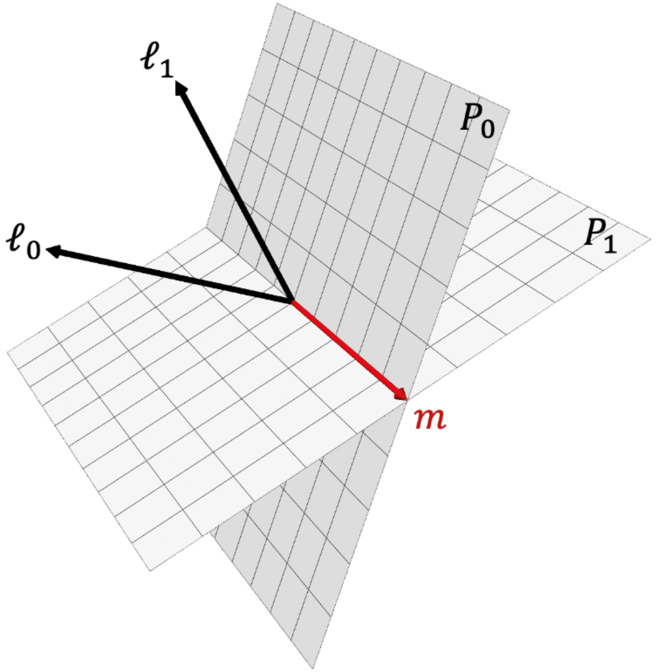
Alfréd Rényi

William G. Brown

Basic idea (geometry and a little Galois theory)

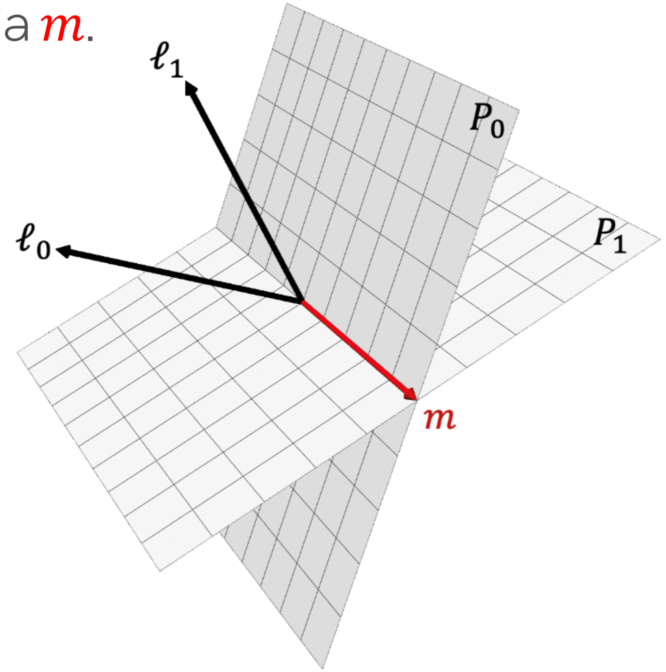
Basic idea (geometry and a little Galois theory)

- If $l_0 \neq l_1$ are any two vectors, there is a vector m perpendicular to both.
  - (It's the cross-product.)



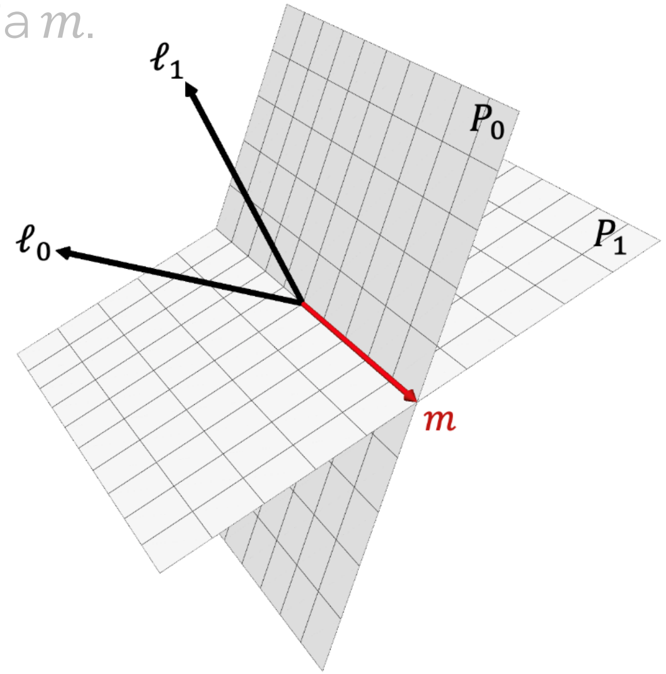
Basic idea (geometry and a little Galois theory)

- If $l_0 \neq l_1$ are any two vectors, there is a vector m perpendicular to both.
  - (It's the cross-product.)
- What if we constructed a graph with edges expressing dot-product perpendicularity?
  - (l_0, m) and (m, l_1) are edges in the graph, so you can get from l_0 to l_1 via m.
  - So this graph has diameter $\mathbf{2}$.



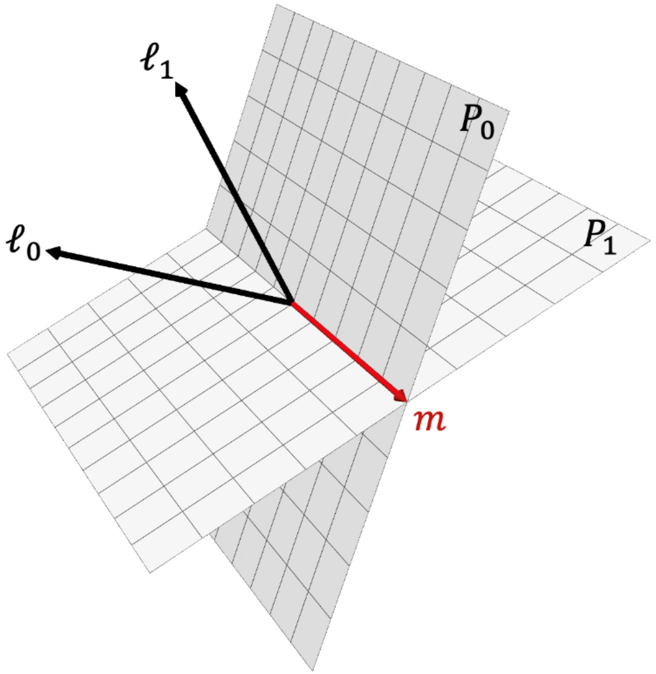
Basic idea (geometry and a little Galois theory)

- If $l_0 \neq l_1$ are any two vectors, there is a vector m perpendicular to both.
  - (It's the cross-product.)
- What if we constructed a graph with edges expressing dot-product perpendicularity?
  - (l_0, m) and (m, l_1) are edges in the graph, so you can get from l_0 to l_1 via m.
  - So this graph has **diameter 2**.
- Use non-0 vectors from $\mathbb{F}_q^3$ whose first non-0 entry is 1:
  - Fact: each is perpendicular to q+1 vectors, so degree is q+1.
  - So the diameter-2 Moore bound is $\boldsymbol{q^2 + 2q + 2}$.
  - Fact: number of nodes/vectors is $\boldsymbol{q^2 + q + 1}$.



Basic idea (geometry and a little Galois theory)

- If $l_0 \neq l_1$ are any two vectors, there is a vector m perpendicular to both.
  - (It's the cross-product.)
- What if we constructed a graph with edges expressing dot-product perpendicularity?
  - (l_0, m) and (m, l_1) are edges in the graph, so you can get from l_0 to l_1 via m.
  - So this graph has **diameter 2**.
- Use non-0 vectors from $\mathbb{F}_q^3$ whose first non-0 entry is 1:
  - Fact: each is perpendicular to q+1 vectors, so degree is q+1.
  - So the diameter-2 Moore bound is $\boldsymbol{q^2 + 2q + 2}$.
  - Fact: number of nodes/vectors is $\boldsymbol{q^2 + q + 1}$.
- *The number of nodes is pretty close to the Moore bound!*



Summing up ER_q ...

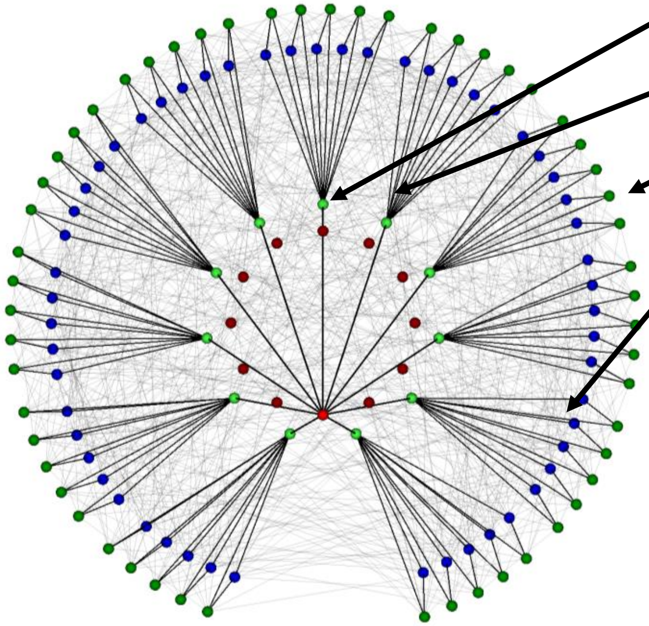
- ER_q has diameter 2.
- There are $q^2 + q + 1$ nodes in ER_q.
- Each non-quadric node has degree $q + 1$
  - So, the graph has degree $q + 1$, and the Moore bound is $(q + 1)^2 + 1$.

- **ER_q asymptotically approaches the Moore bound:**

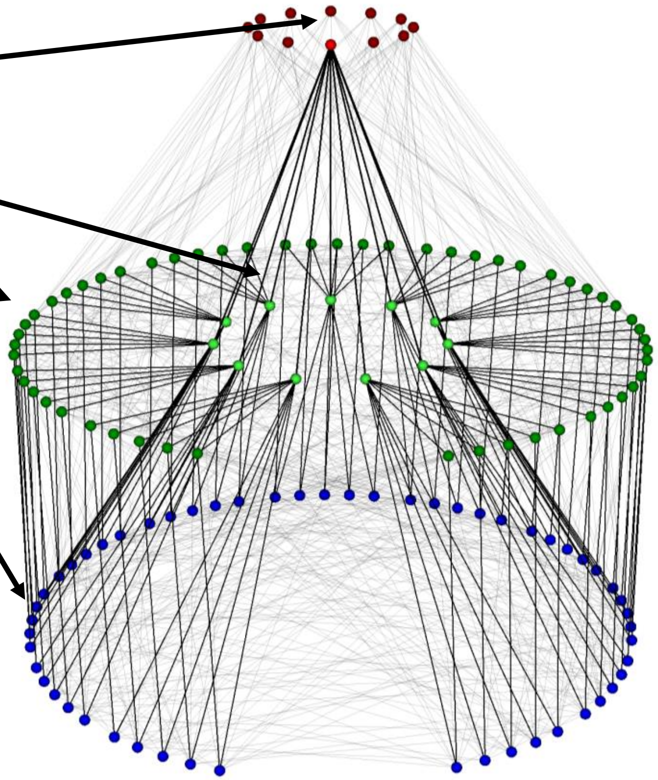
$$\frac{\#\, nodes}{Moore\ bound} = \frac{q^2+q+1}{(q+1)^2+1} \Longrightarrow 1, \text{as } q \Longrightarrow \infty$$

Can ER_q form the basis for laying out a high-radix network?

Yes … Introducing: *PolarFly**



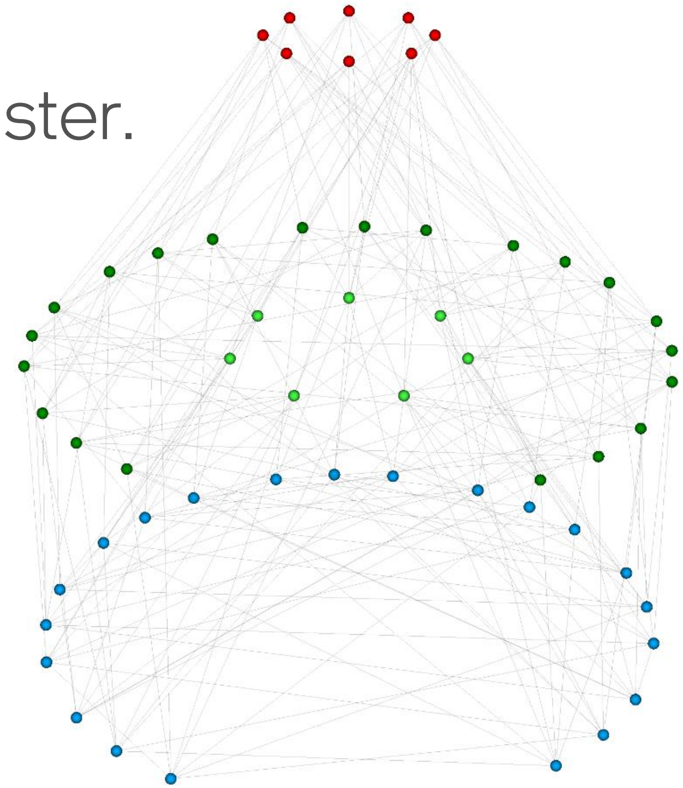
Four types of vertices
1. Quadrics
2. Centers
3. V1
4. V2



* K. Lahotia, M. Besta, **L. Monroe**, K. Isham, P. Iff, T. Hoefler, and **F. Petrini**. "PolarFly: A Cost-Effective and Flexible Low-Diameter Topology". The *International Conference for High Performance Computing, Networking, Storage, and Analysis* (SC22). November 2022. https://arxiv.org/abs/2208.01695.

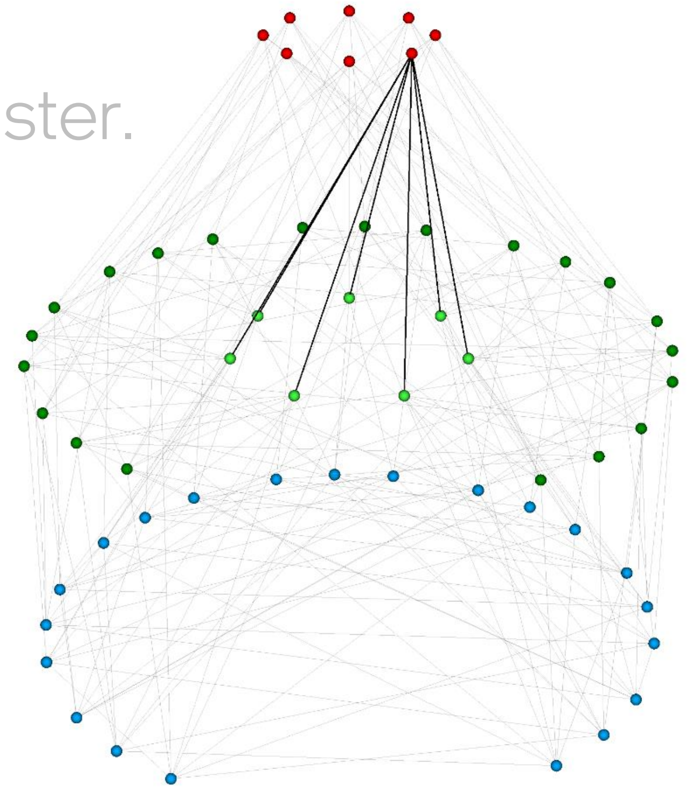
Structure of PolarFly

- All self-perpendicular quadrics (red) form a cluster.



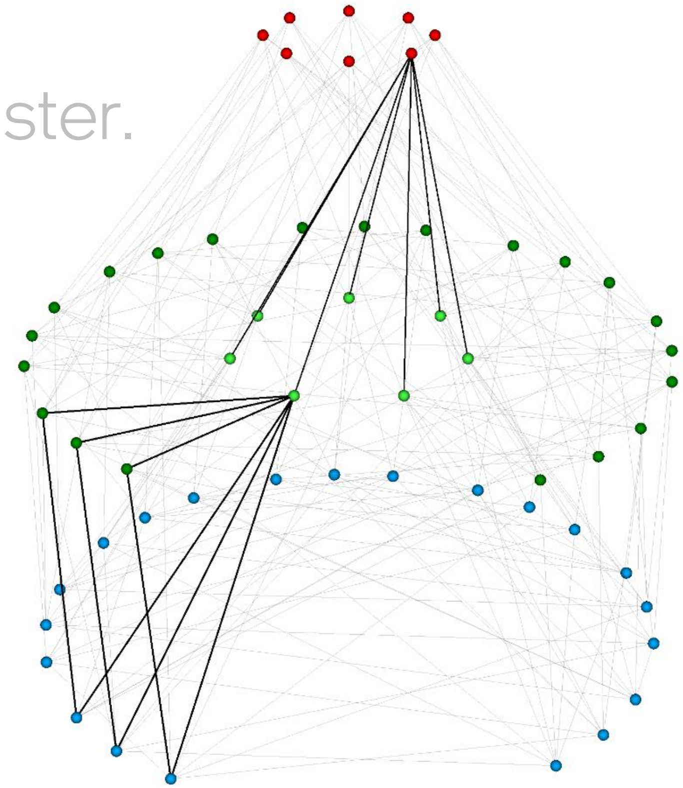
Structure of PolarFly

- All self-perpendicular quadrics (red) form a cluster.
- Pick one as the starter quadric l.
- Take all vectors c perpendicular to l.
  - These are the centers.



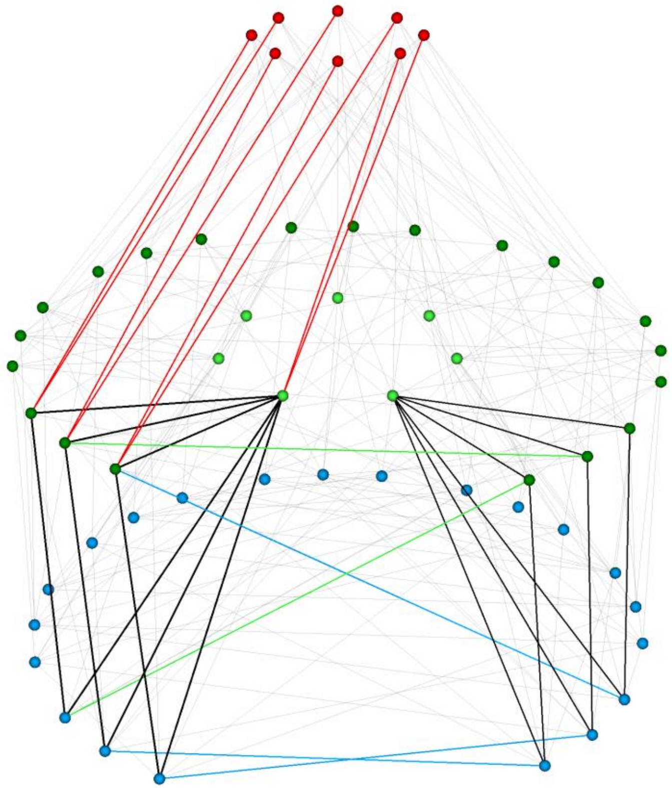
Structure of PolarFly

- All self-perpendicular quadrics (red) form a cluster.
- Pick one as the starter quadric l.
- Take all vectors c perpendicular to l.
  - These are the centers.
- Each center c starts its own cluster:
  - All vectors v perpendicular to c.



Structure of PolarFly

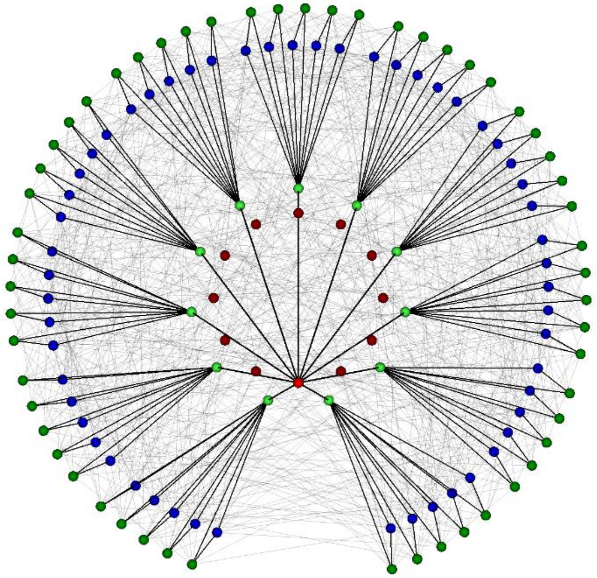
- All self-perpendicular quadrics (red) form a
- Pick one as the starter quadric l.
- Take all vectors c perpendicular to l.
  - These are the centers.
- Each center c starts its own cluster:
  - All vectors v perpendicular to c.



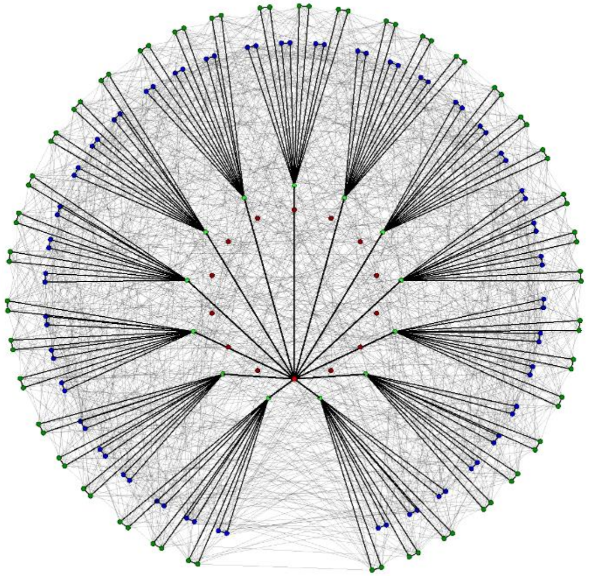
- Each non-quadric cluster has $q+1$ connections to the quadric cluster.
- Each non-quadric cluster has $q-2$ connections to non-quadric clusters.

Structure: Triangles internal to non-quadric clusters

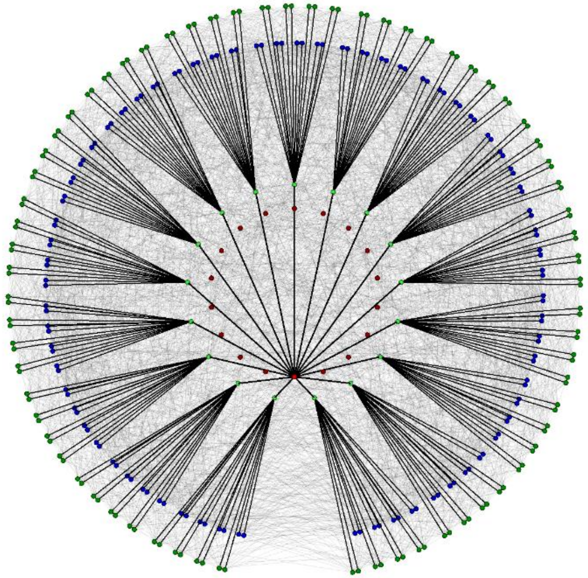
- There are q non-quadric clusters.
- Each non-quadric cluster is a fan-out of $\frac{q-1}{2}$ triangles.



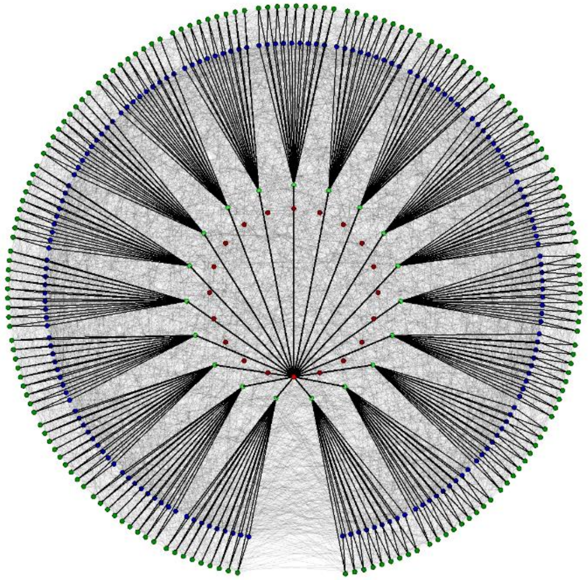
$q = 11$



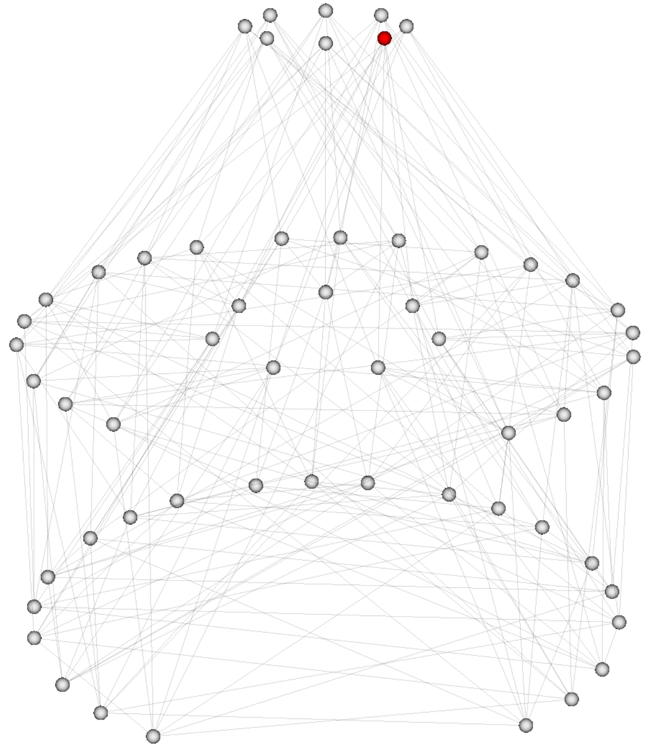
$q = 13$



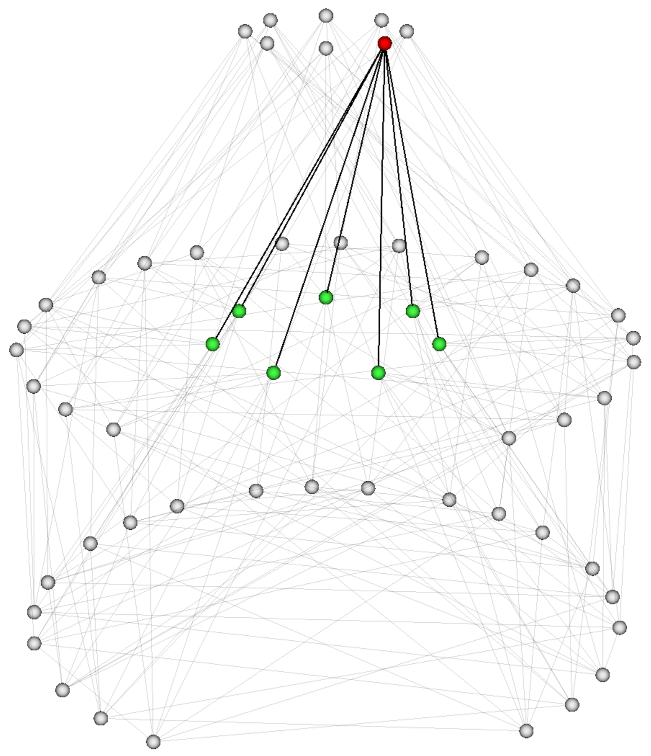
$q = 17$



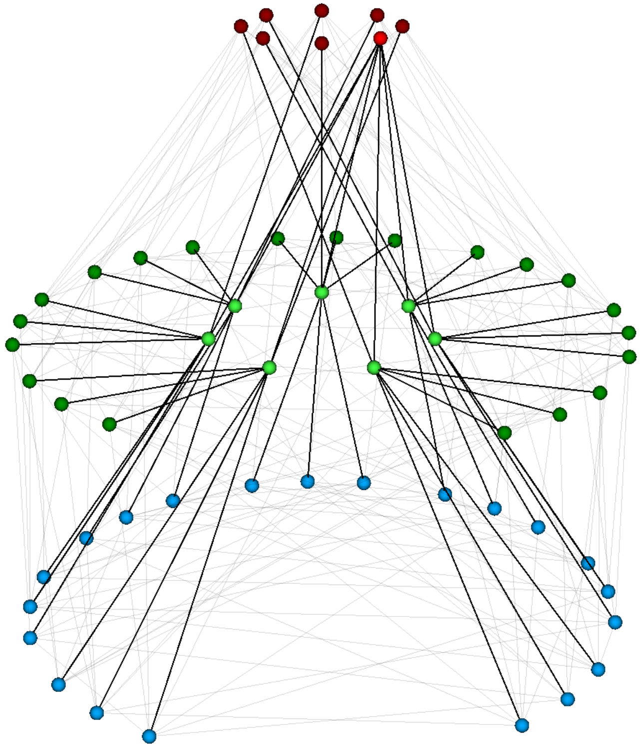
$q = 19$



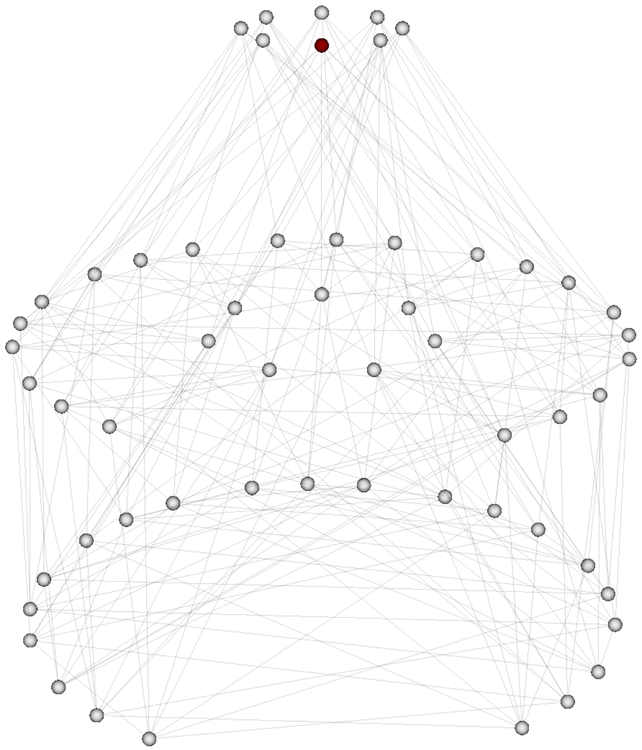
0 hops from starter quadric.



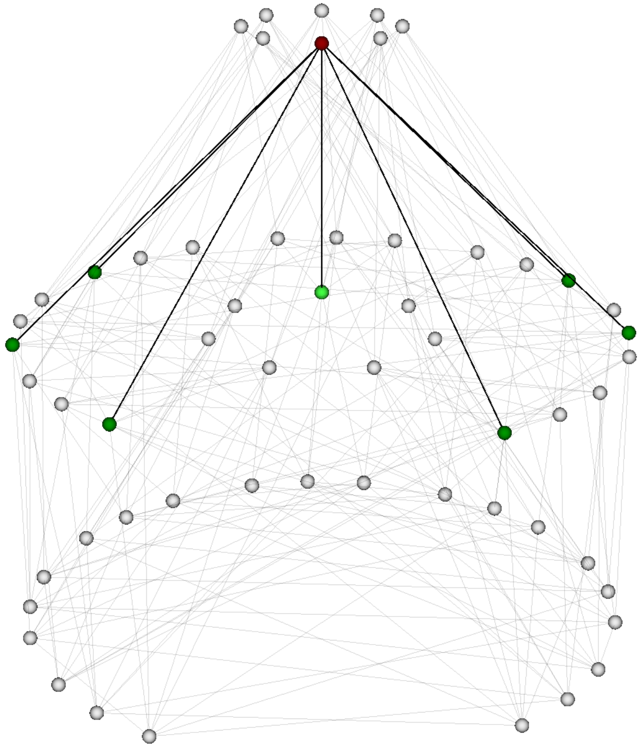
1 hop from starter quadric.



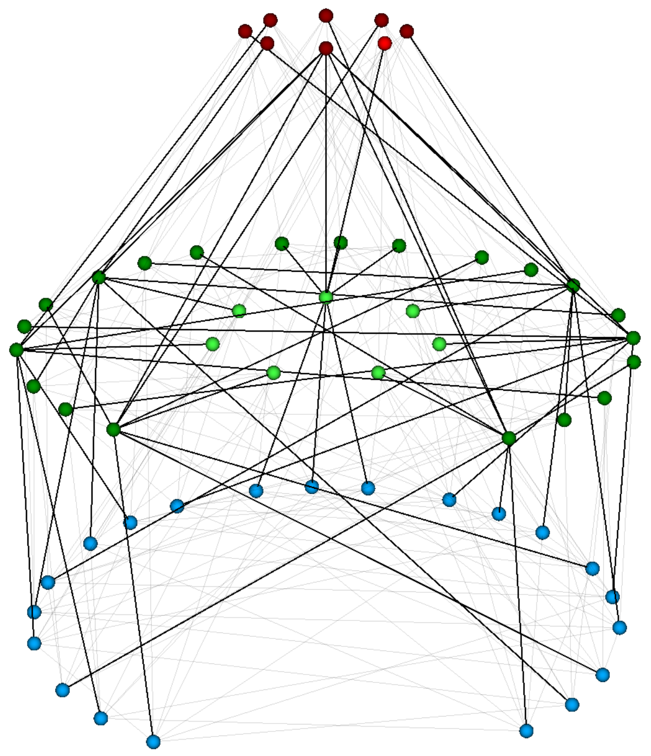
2 hops from starter quadric.



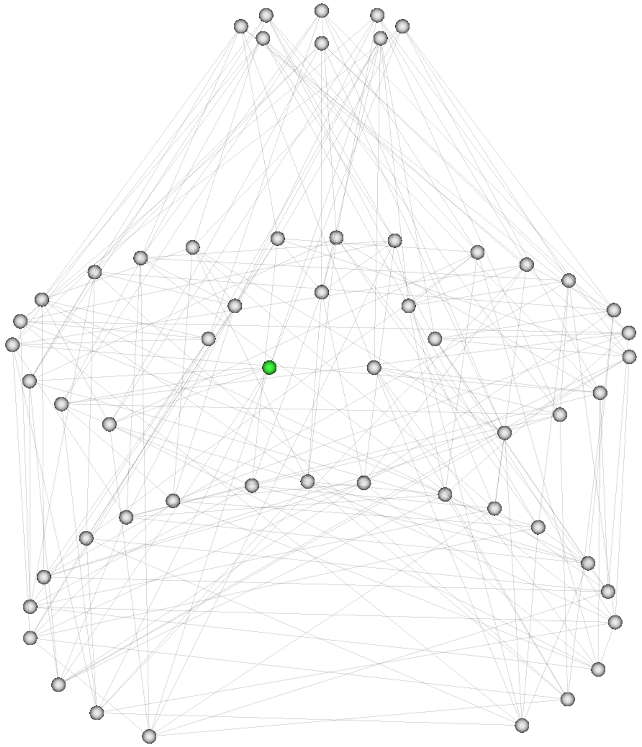
0 hops from another quadric.



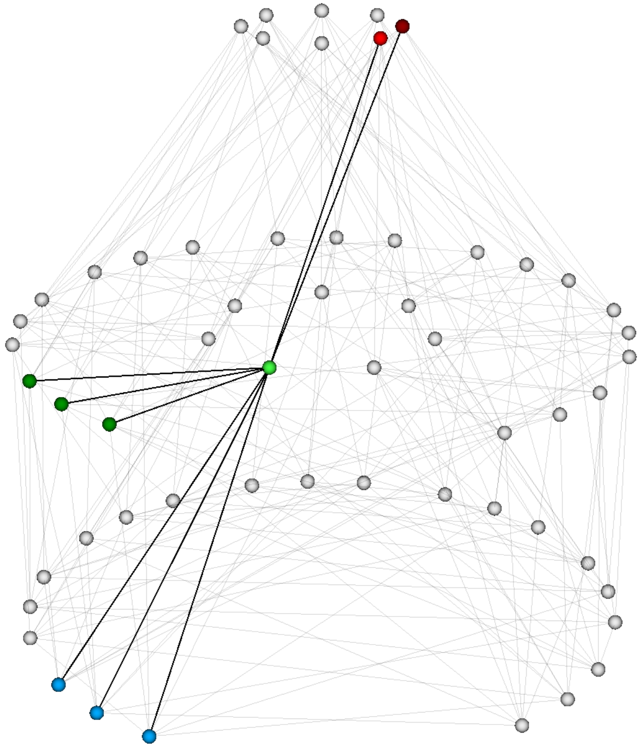
1 hop from another quadric.



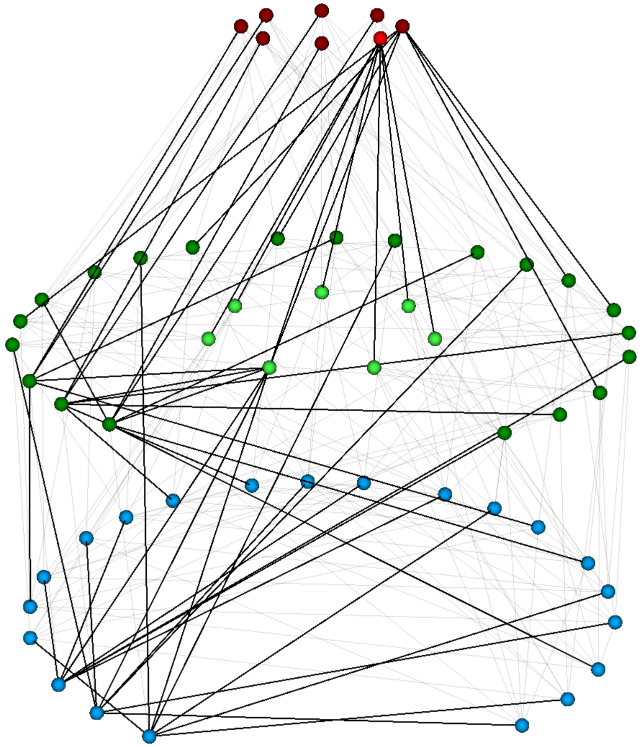
2 hops from another quadric.



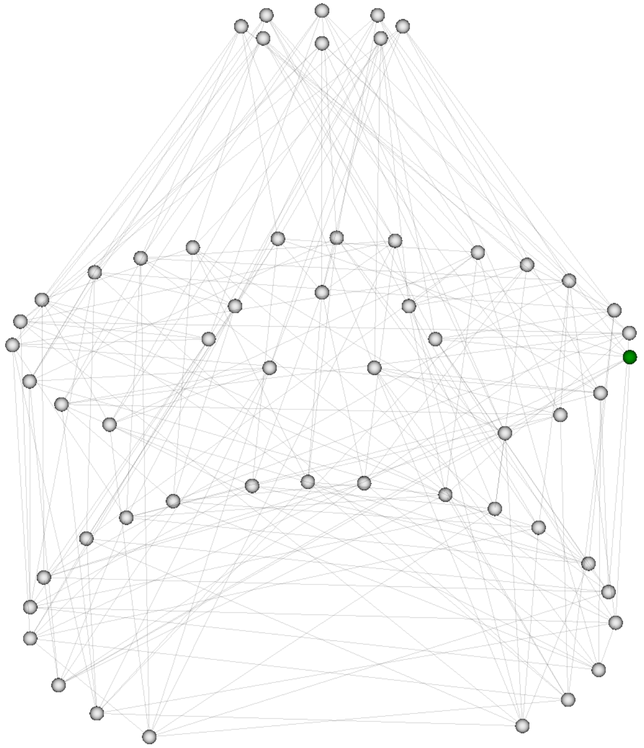
0 hops from a center element.



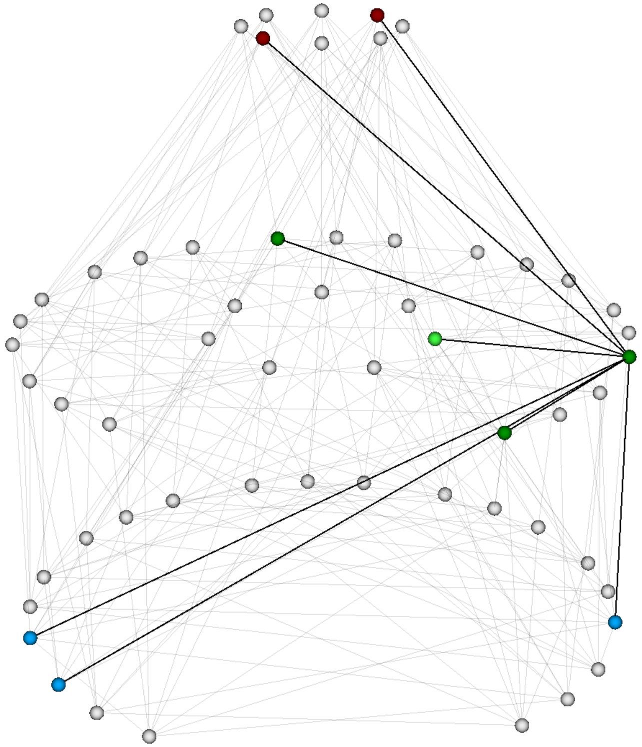
1 hop from a center element.



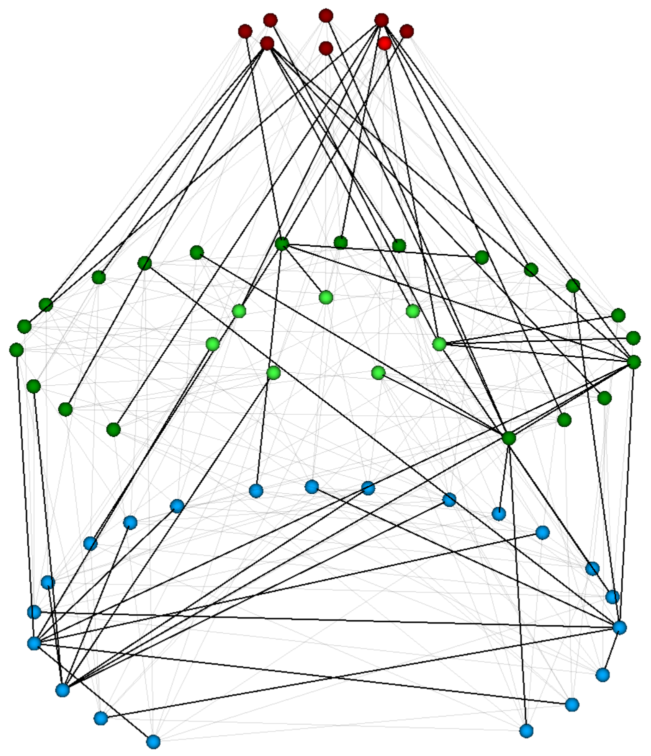
2 hops from a center element.



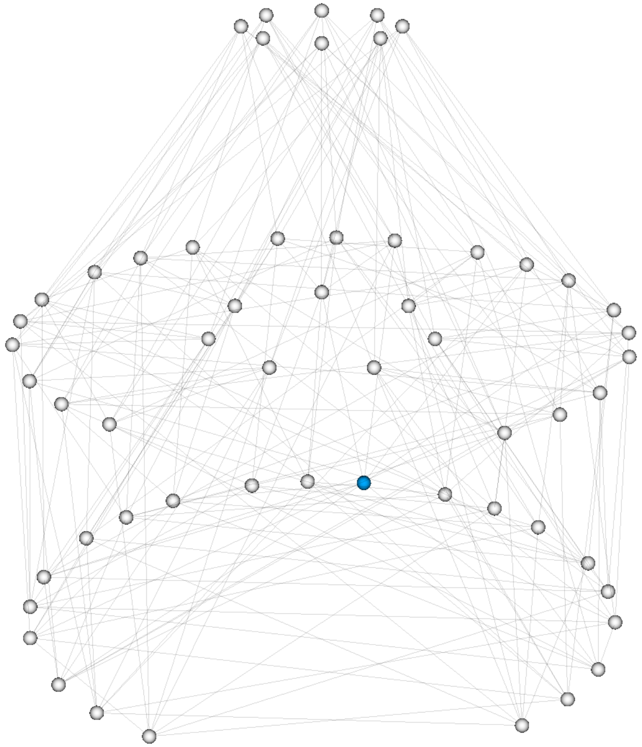
0 hops from a V1 element.



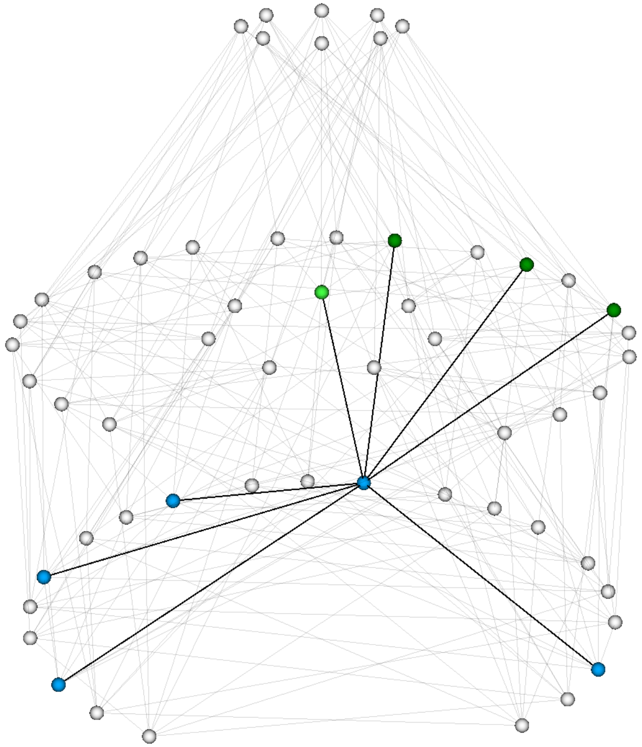
1 hop from a V1 element.



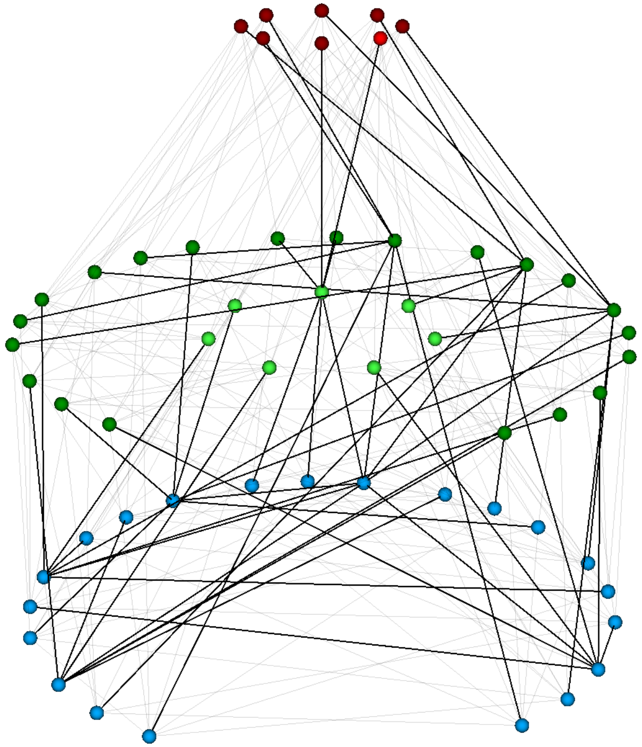
2 hops from a V1 element.



0 hops from a V2 element.



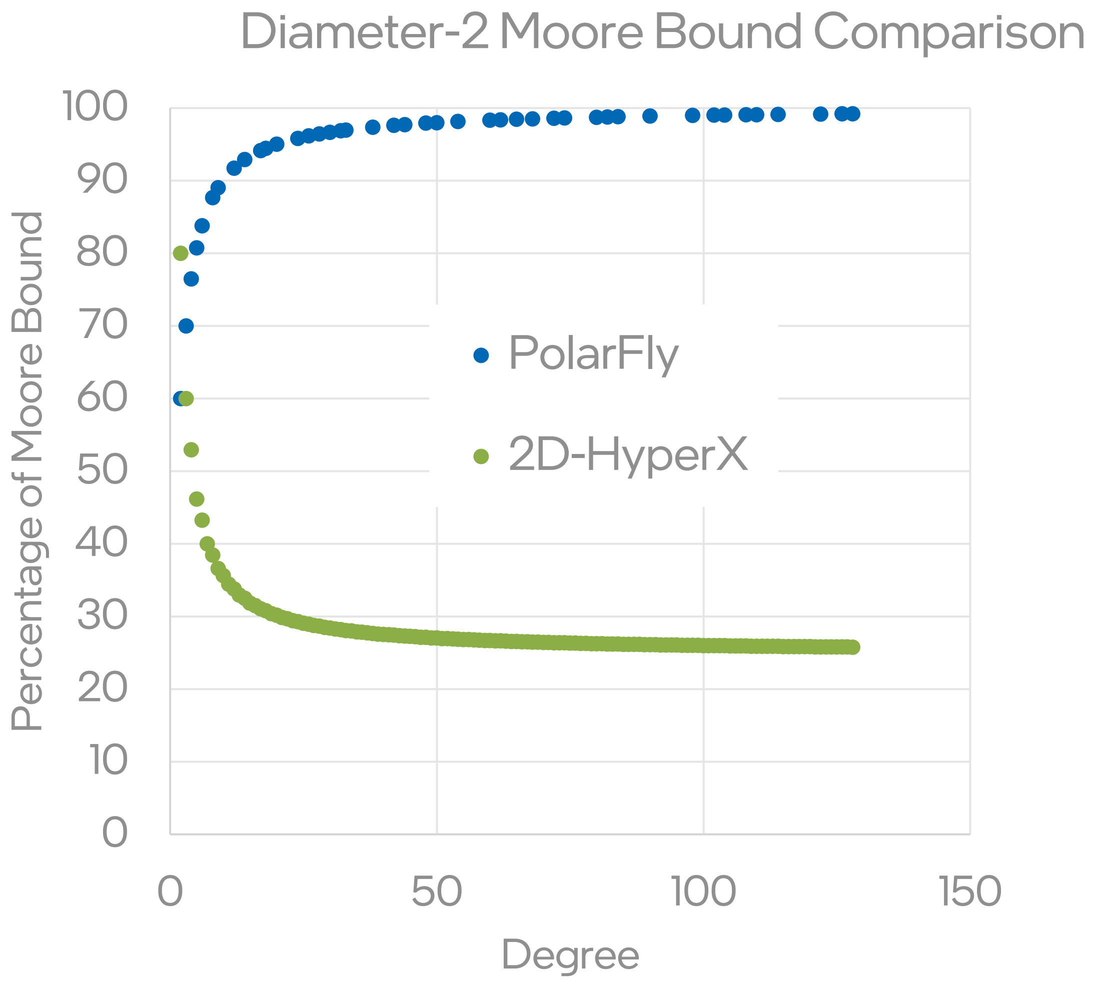
1 hop from a V2 element.



2 hops from a V2 element.

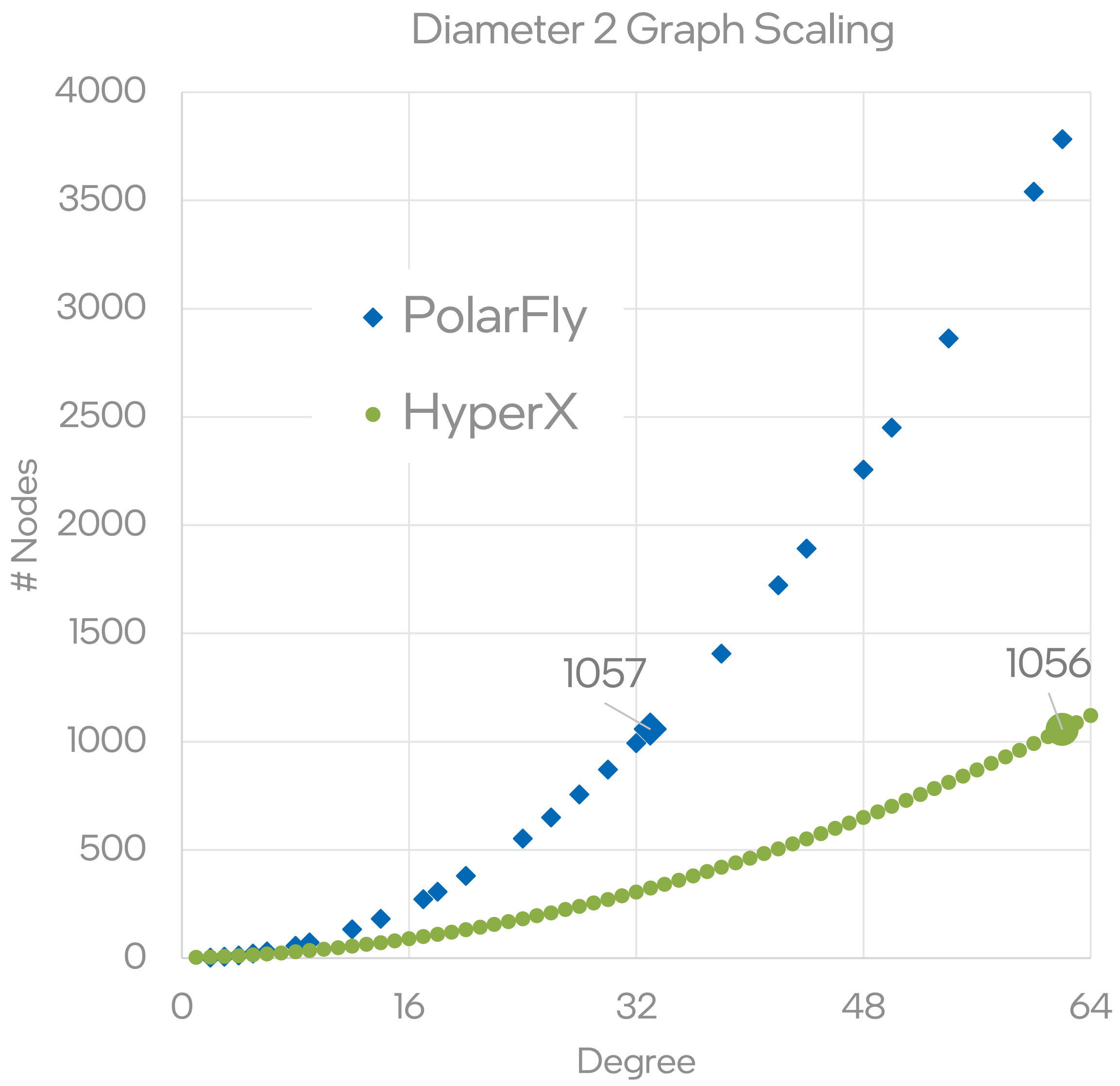
PolarFly: Scalability

- Provably optimal scale for a given radix
- Reaches Moore Bound asymptotically
- More flexible and scalable than prior-art



PolarFly: Scalability … continued

- HyperX requires radix 62 to connect 1,056 nodes
  - 32,736 cables and optical IO modules
- PolarFly can achieve the same scale with radix 33
  - Only 17,424 cables and optical IO modules
- Cost-savings & better performance



To sum up …

- AI is redefining not just what compute can do for us, but also how we do compute
- Demand for compute is scaling faster than we can meet
- Memory and networking growth are falling behind, creating an ever-larger gap with compute
- Significant cost of sustaining AI compute scaling lies in meeting its energy cost
- Increasing fraction of energy is spent moving data, not computing on data
- High-radix optical networks have the potential to significantly address network latency, energy and cost.
- Polar Fly offers a promising basis for building a high-radix, diameter-2 network that can scale-up to thousands of GPUs.
  - Diameter-3 is research-in-progress

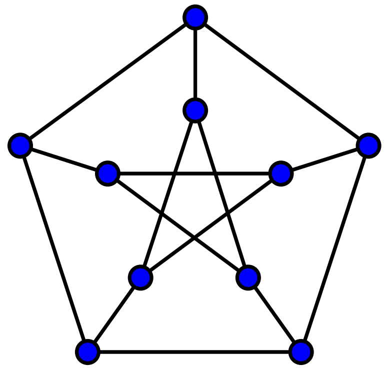
Thank you for your time!

- Questions?

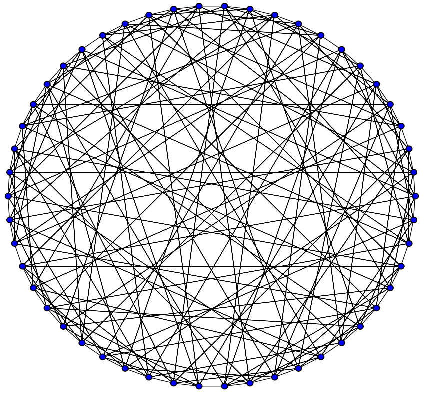
Back up

Are there graphs that meet the Moore bound?

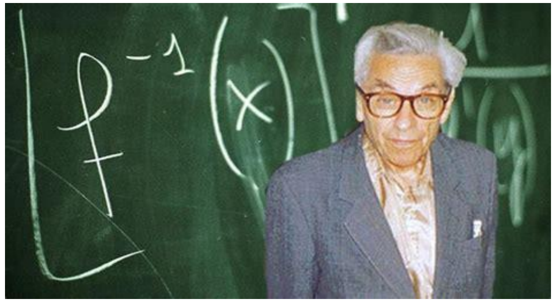
- Yes, but not very many.
  - The family of complete graphs K_n
  - Diameter 2 with degrees 2, 3, 7 and maybe 57
- **How about asymptotically? Yes!**
  - **The Erdős-Rényi (ER) polarity graphs do.**



Petersen graph: degree 3, 10 routers



Hoffman-Singleton graph: degree 7, 50 routers



Paul Erdős

Alfréd Rényi