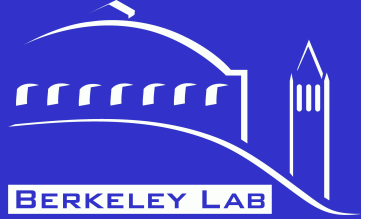
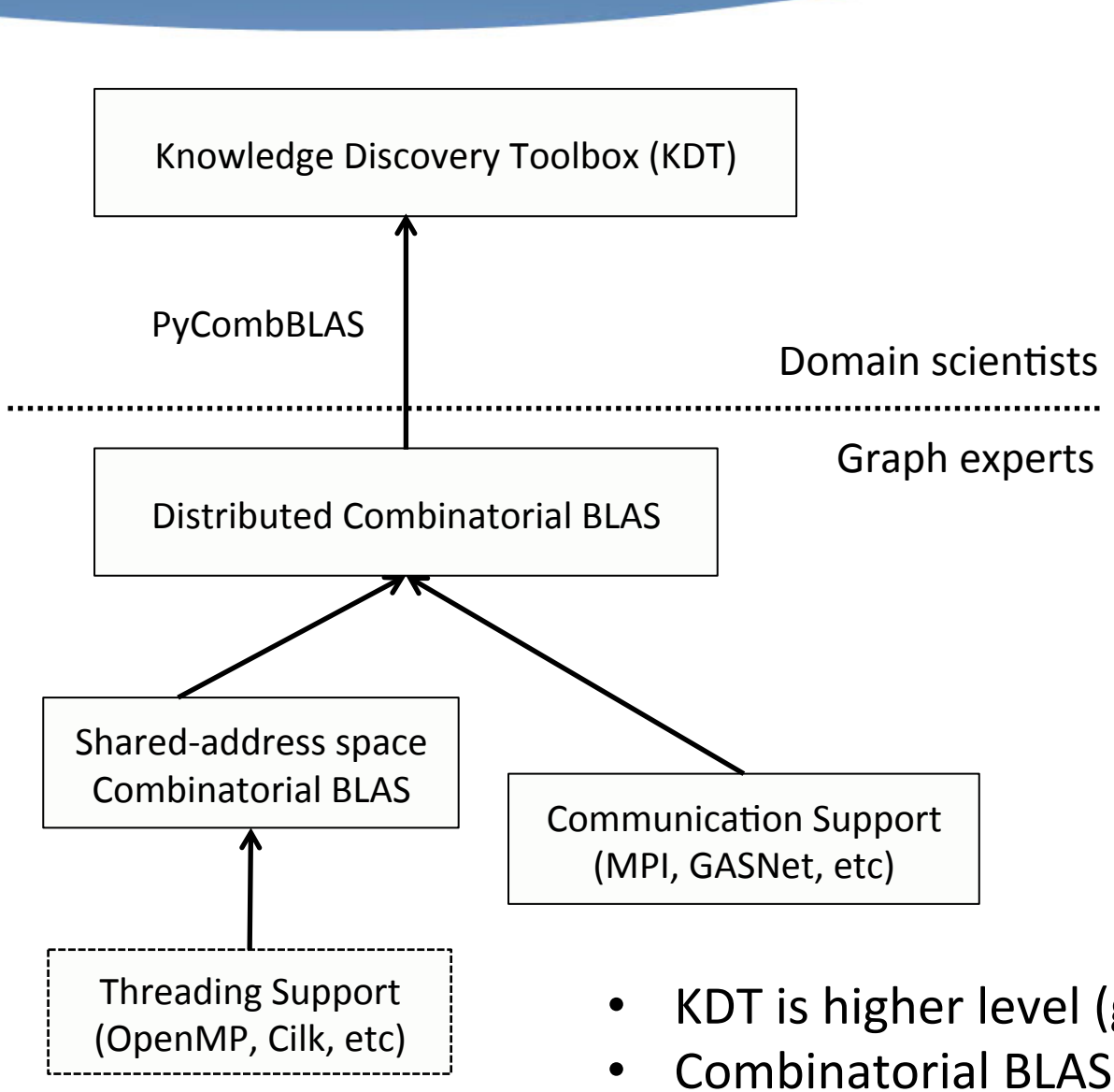


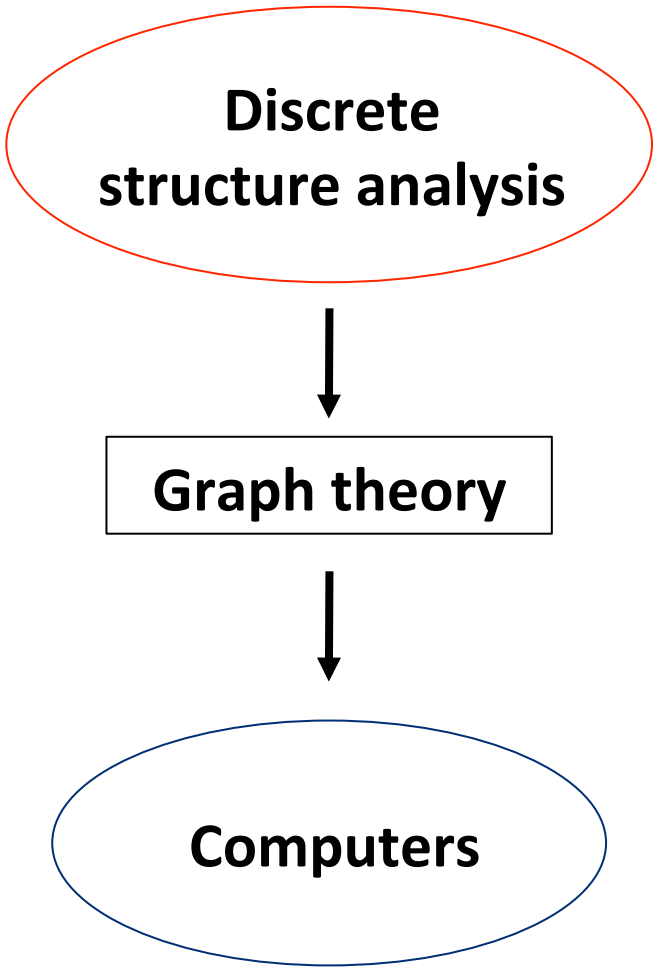
Parallel Graph Libraries: Where do we go from here?

Aydın Buluç
Lawrence Berkeley National Laboratory

KDT Spring Mind Meld
March 5, 2012

(Proposed) Software for Graph Analysis



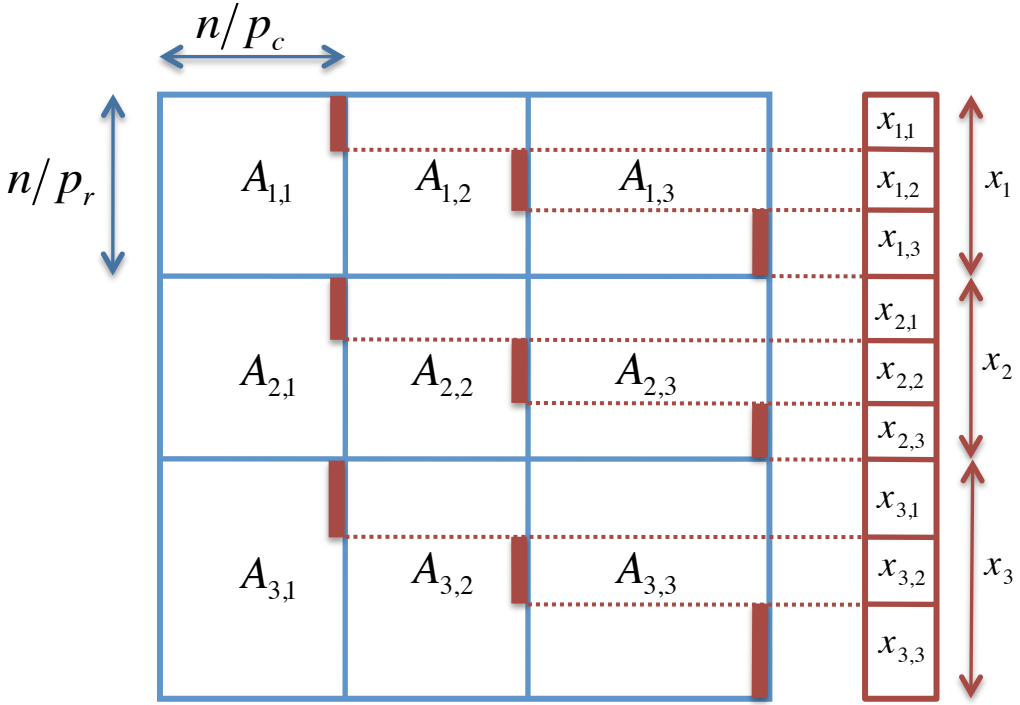


- KDT is higher level (graph abstractions)
- Combinatorial BLAS is for performance

Outline

- The bottleneck: Communication
- The problem with graph partitioners
- Architectural evolution
- Data diversity: Graph characteristics
- Algorithmic evolution: Beat the worst case
- Functionality evolution: Most important kernels
- Peripherals: Database/visualization integration

2D layout for sparse matrices & vectors

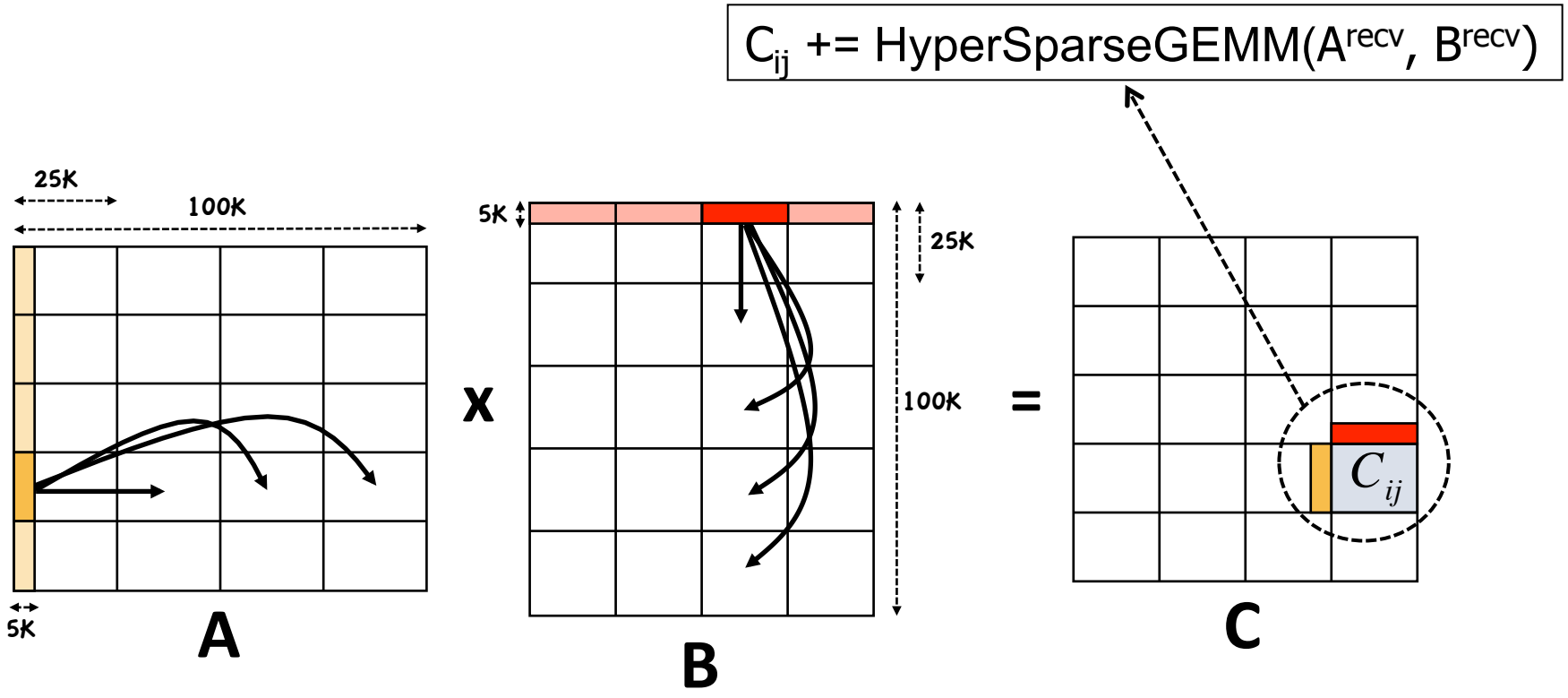


Matrix/vector distributions, interleaved on each other.

Default distribution in **Combinatorial BLAS**.

- 2D matrix layout wins over 1D with large core counts and with limited bandwidth/compute
- 2D vector layout sometimes important for load balance
- Scalable with increasing number of processes

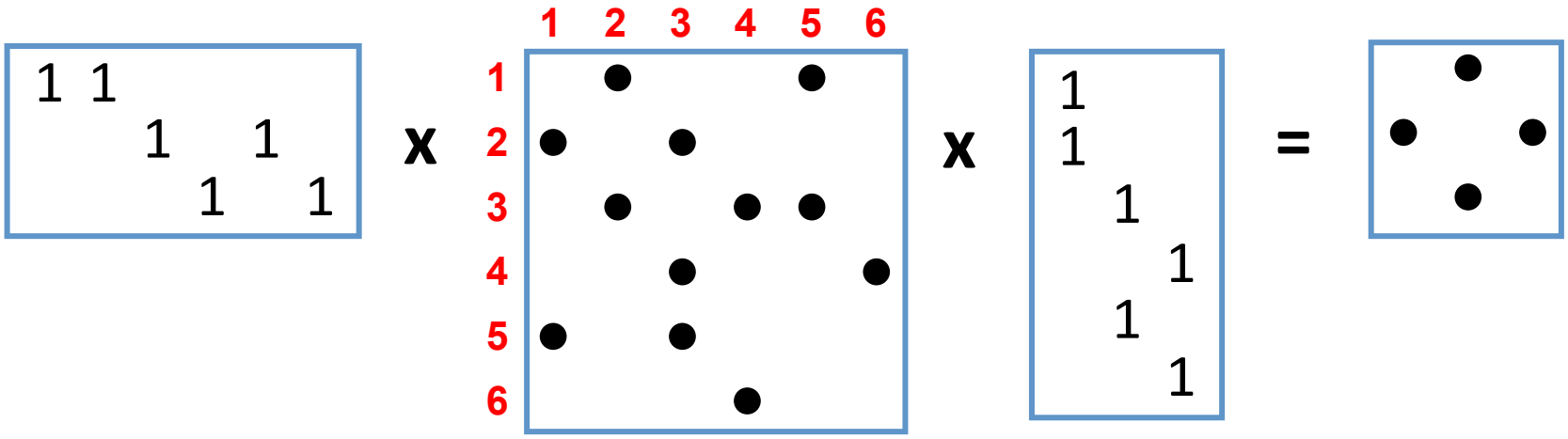
2D algorithm: Sparse SUMMA

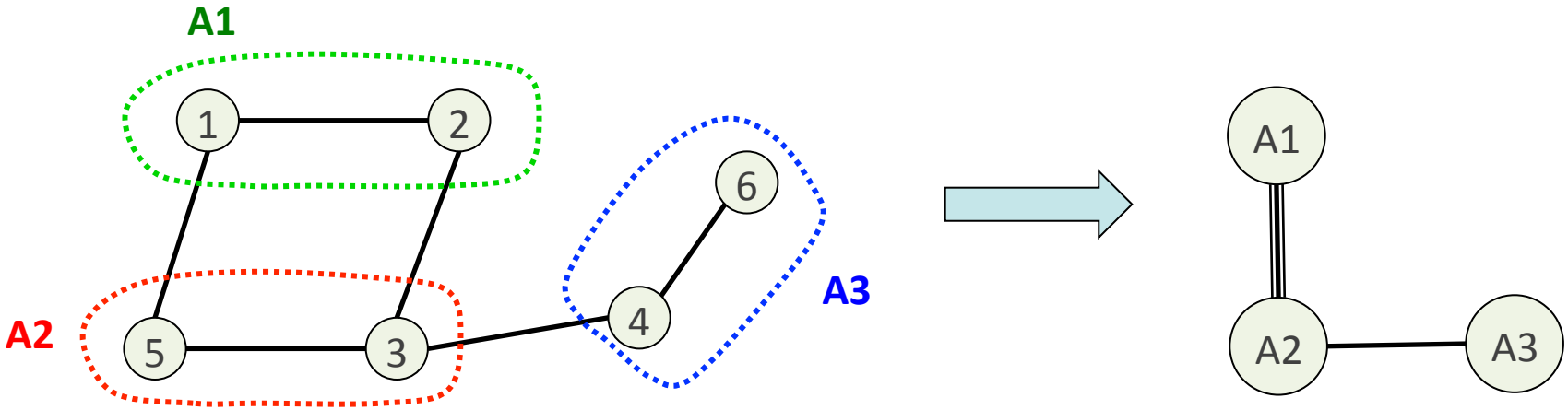


Based on dense SUMMA

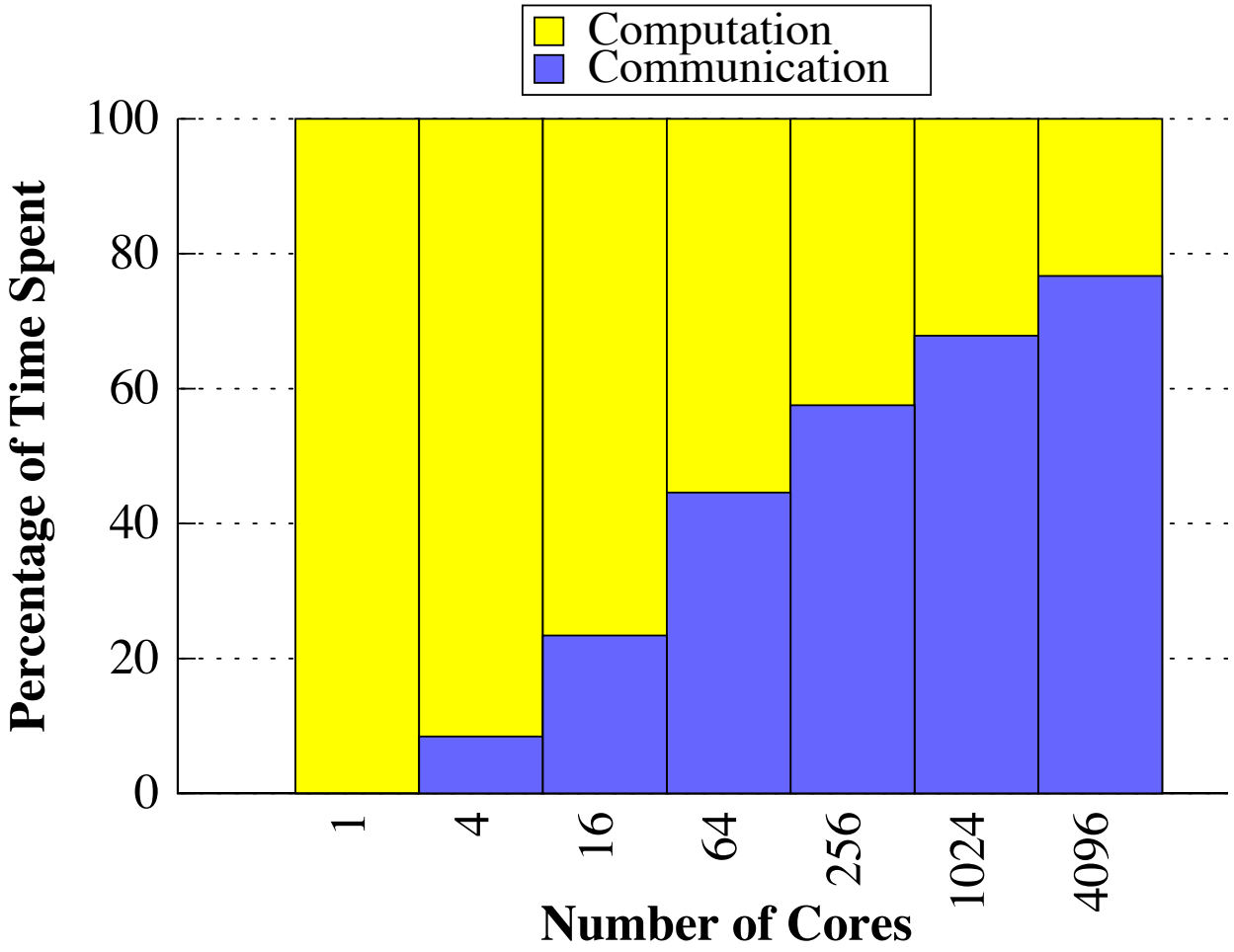
General implementation that handles rectangular matrices

Multiplication with the restriction operator



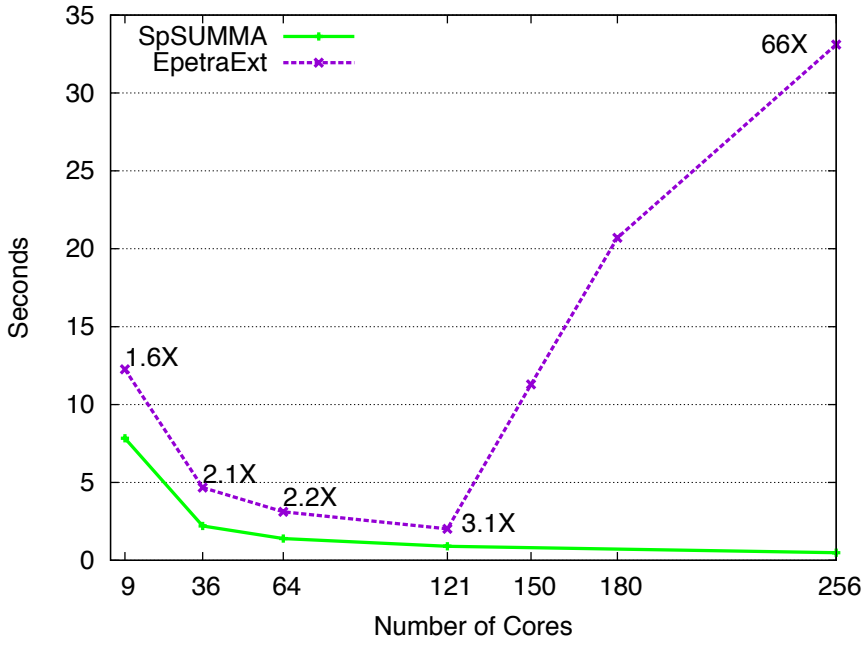


The need to reduce communication

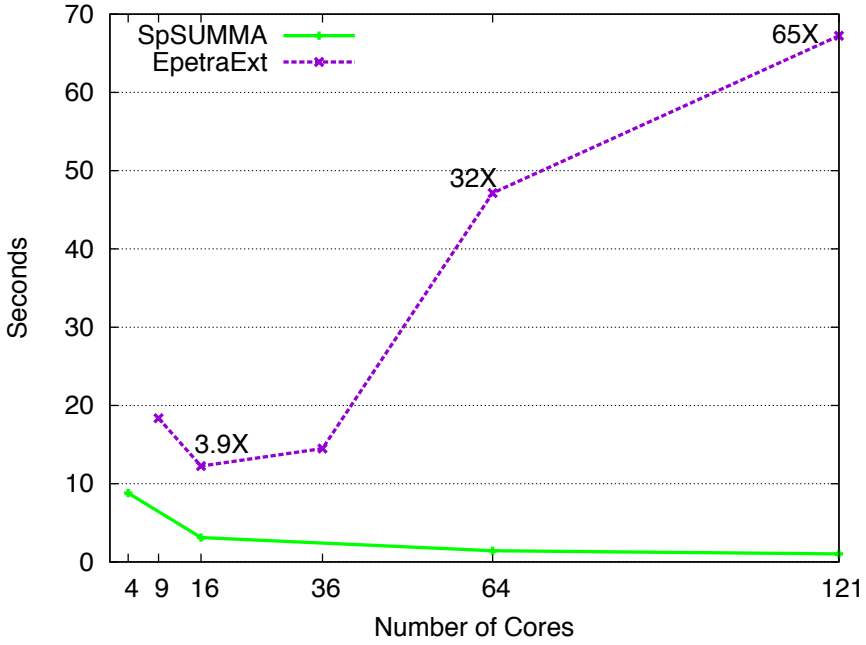


- Normalized communication/computation breakdown
- Scale 23 R-MAT times restriction operator of order 4

Comparison of SpGEMM implementations



(a) R-MAT × R-MAT product (scale 21).

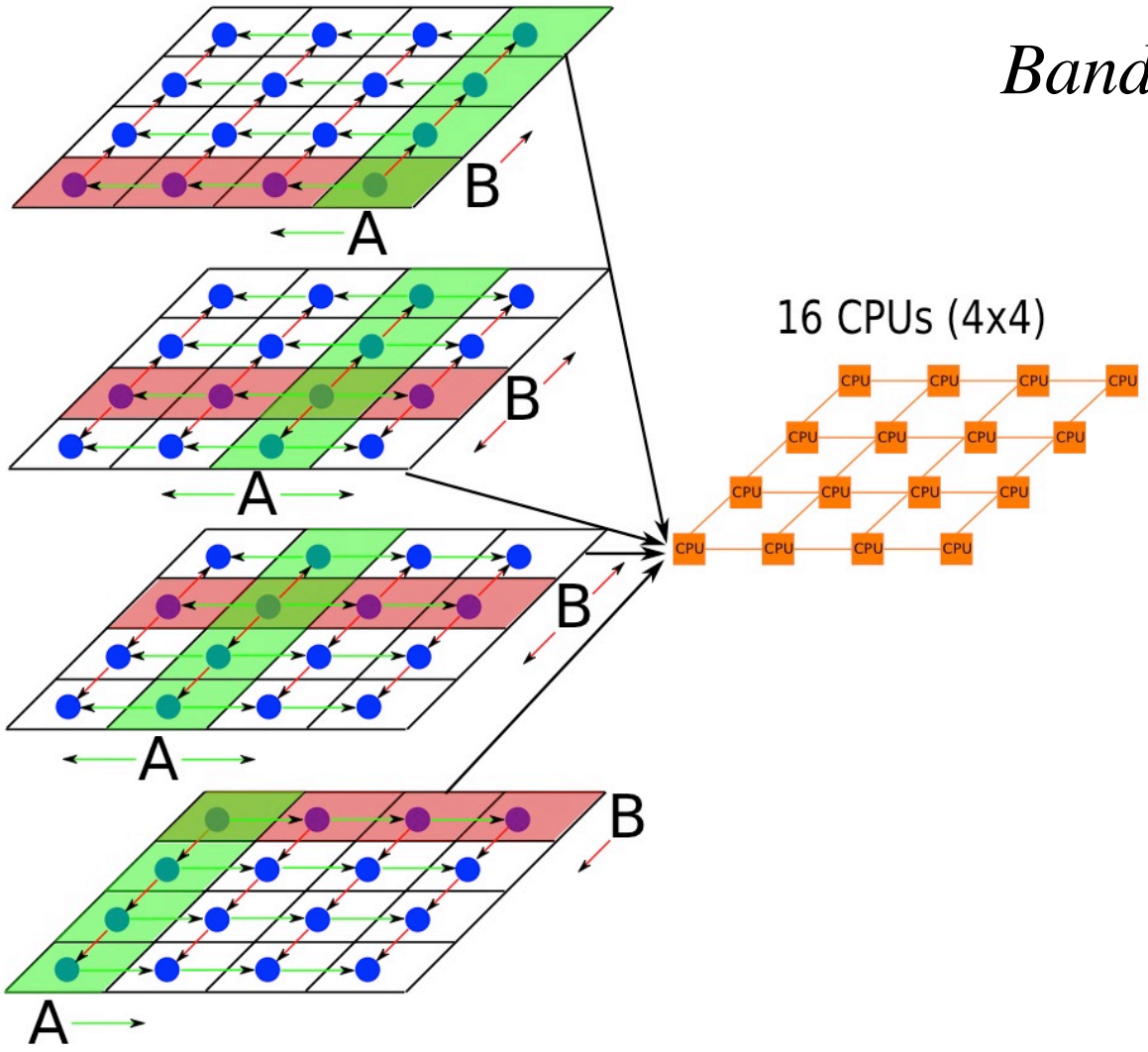


(b) Multiplication of an R-MAT matrix of scale 23 with the restriction operator of order 8.

SpSUMMA = 2-D data layout (Combinatorial BLAS)

EpetraExt = 1-D data layout (Trilinos)

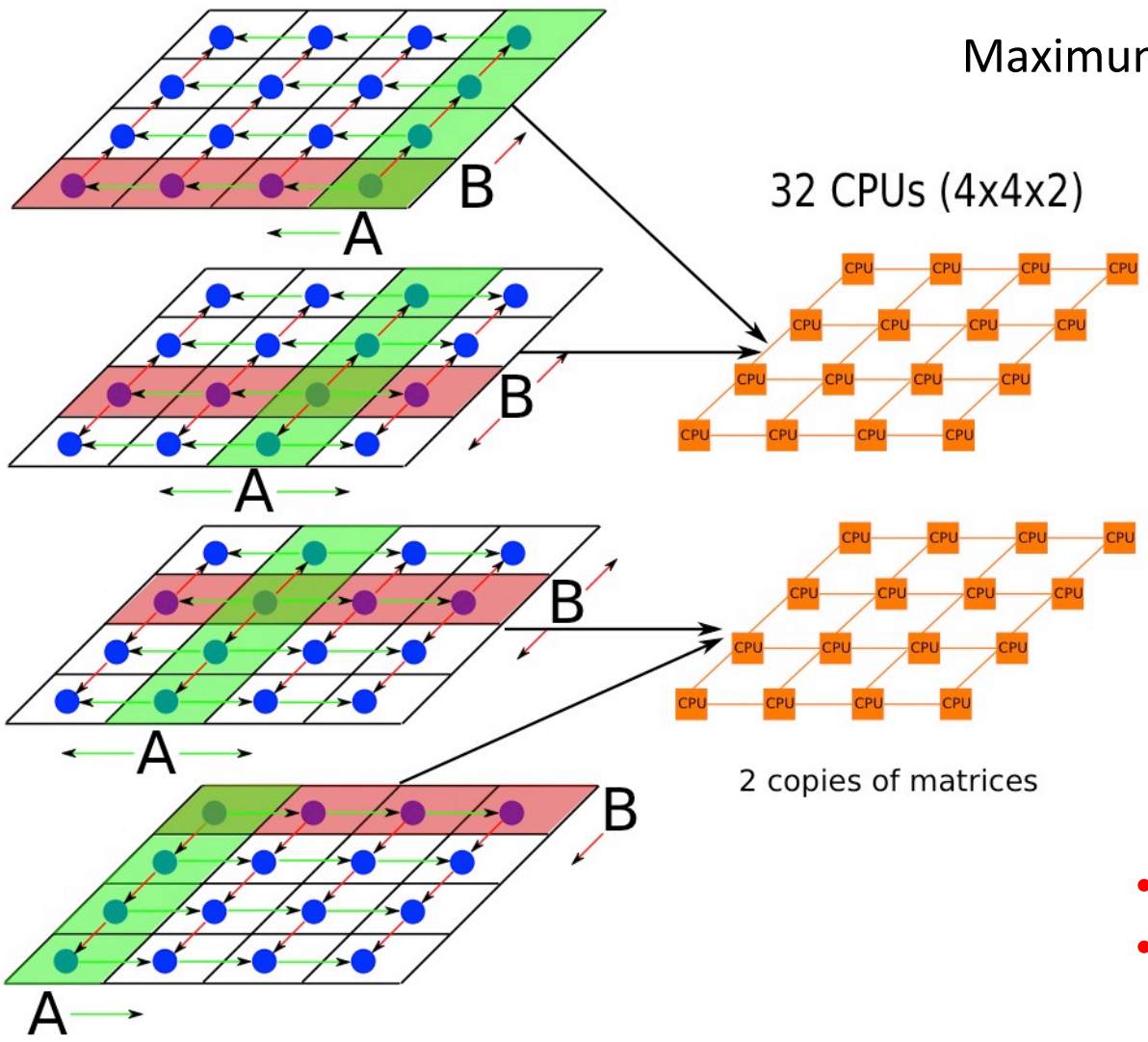
Remember the 2D algorithm



$$Bandwidth = \Theta(\frac{dn}{\sqrt{p}})$$

Generalize SUMMA to 2.5D

[Ballard, B., Demmel, Grigori, Schwartz]

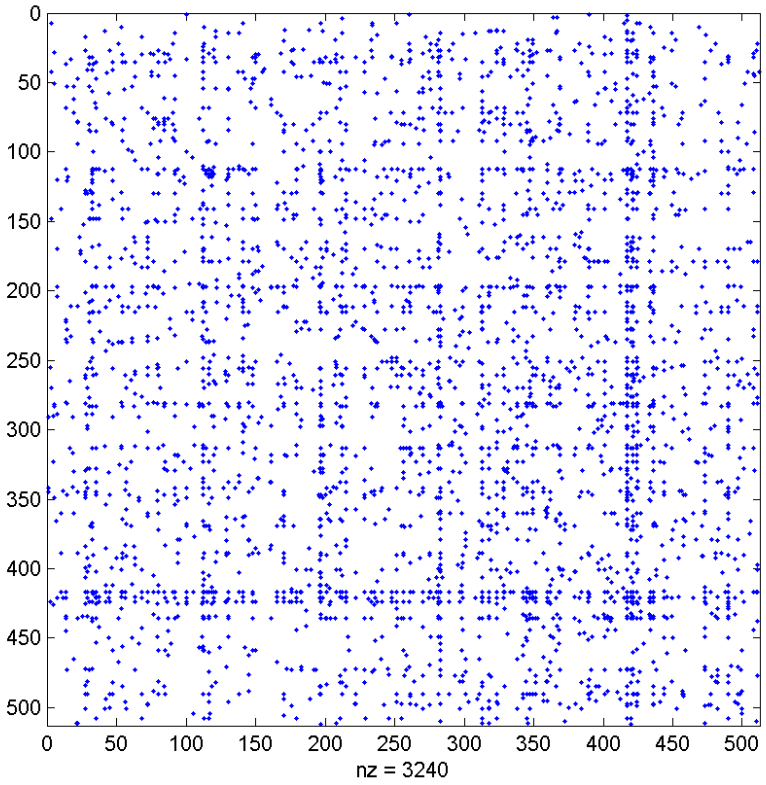


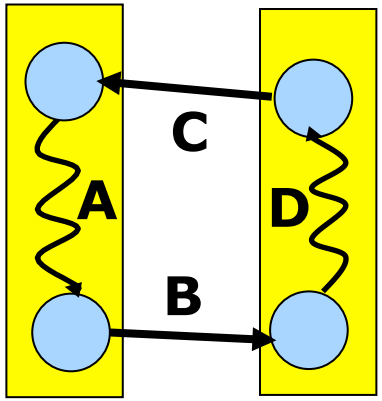
Maximum replicas: $c \leq \frac{\sqrt[3]{p}}{d^{2/3}}$

$$Bandwidth: \Theta\left(\frac{d^{4/3}n}{p^{2/3}}\right)$$

- **Better scaling with p**
- **Worse with d**

Recursive all-pairs shortest paths





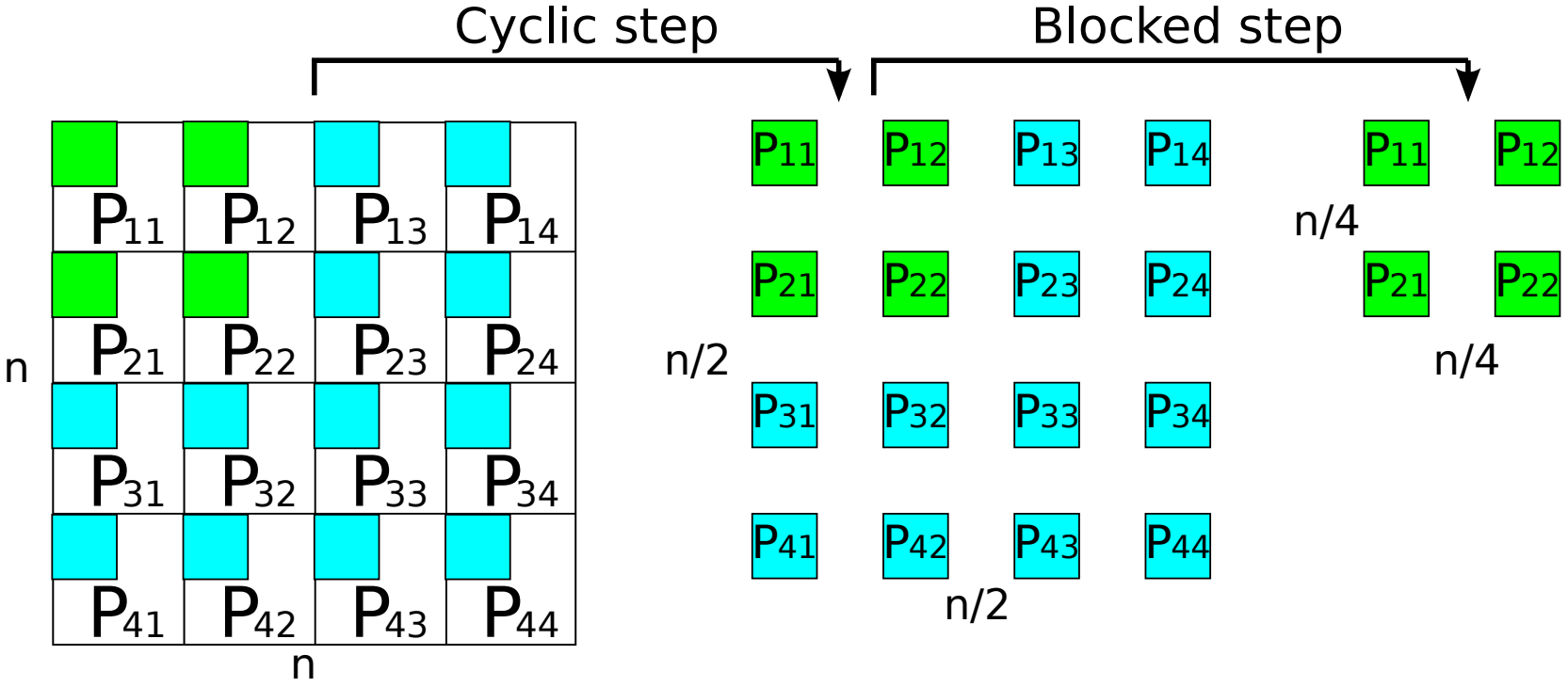
A	B
C	D

+ is "min", **×** is "add"

```
A = A*;     % recursive call
B = AB;  C = CA;
D = D + CB;
D = D*;     % recursive call
B = BD;  C = DC;
A = A + BC;
```

Novel 2.5D APSP algorithm

[Solomonik, B., Demmel; 2012]



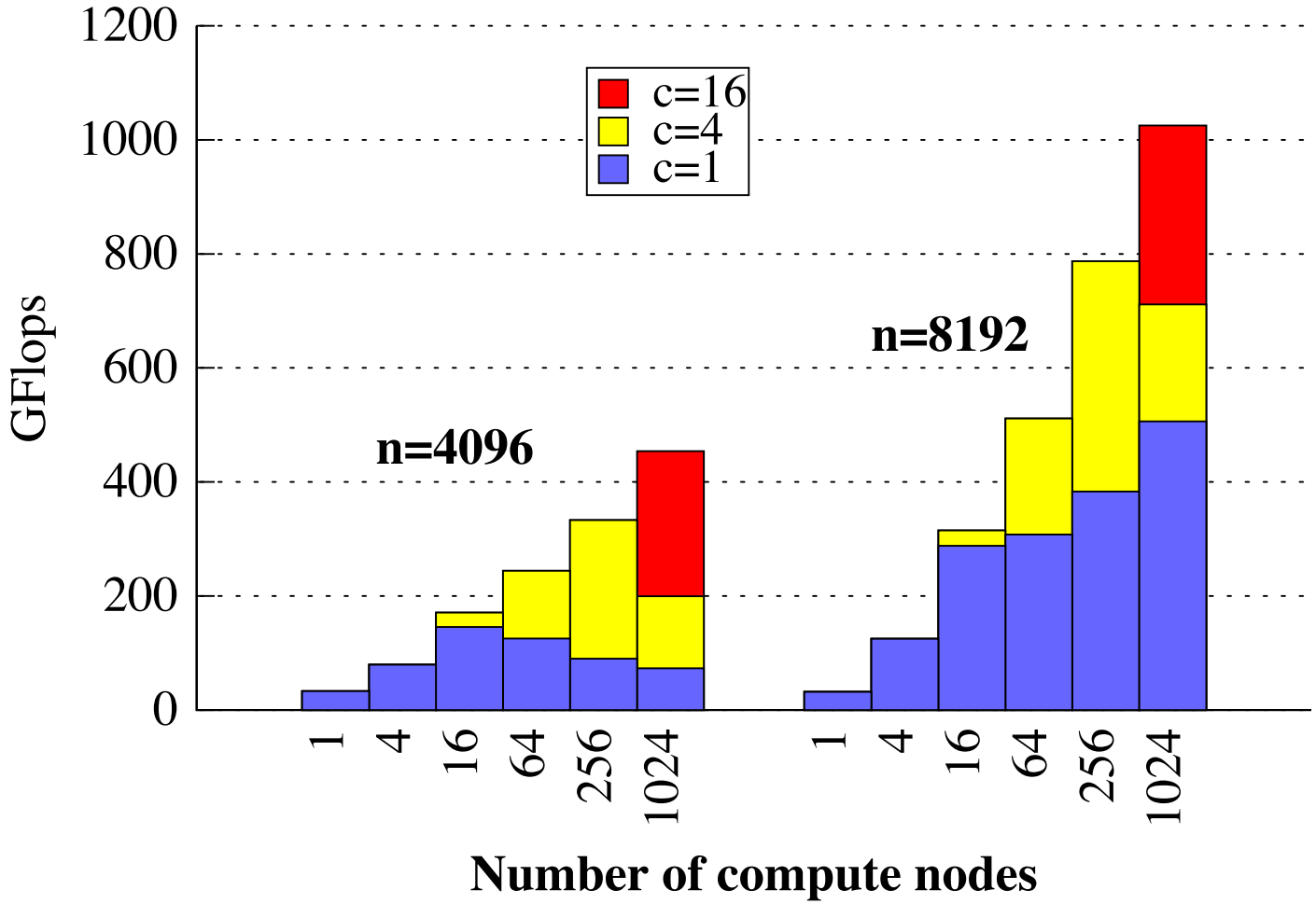
Bandwidth: $W_{\text{bc-2.5D}}(n,p) = O(n^2/\sqrt{cp})$

Latency: $S_{\text{bc-2.5D}}(p) = O\left(\sqrt{cp}\log^2(p)\right)$

c: number of replicas

Optimal for any memory size !

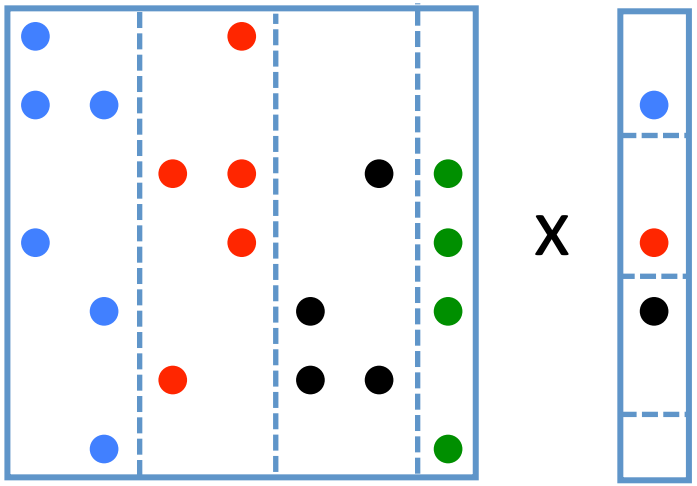
Novel 2.5D APSP algorithm

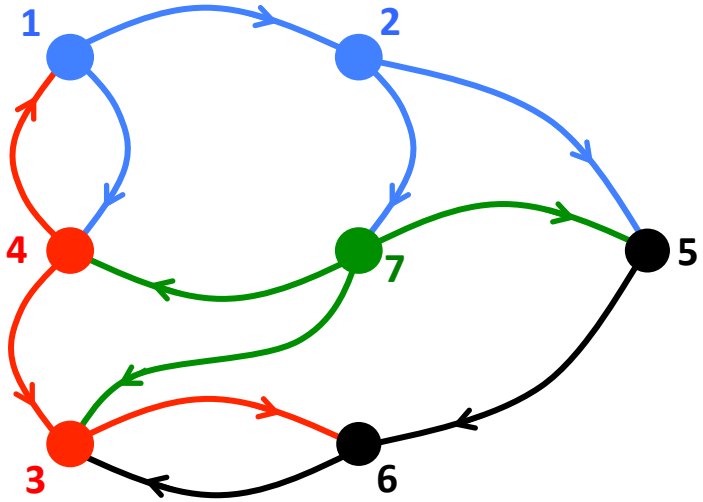


Outline

- The bottleneck: Communication
- The problem with graph partitioners
- Architectural evolution
- Data diversity: Graph characteristics
- Algorithmic evolution: Beat the worst case
- Functionality evolution: Most important kernels
- Peripherals: Database/visualization integration

1D parallel BFS algorithm



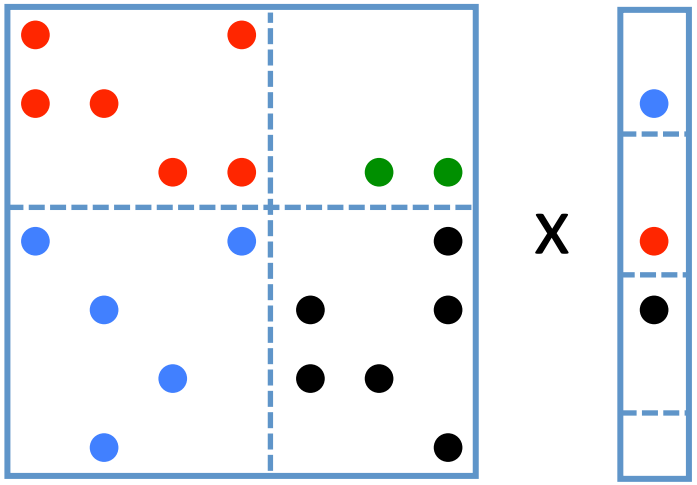


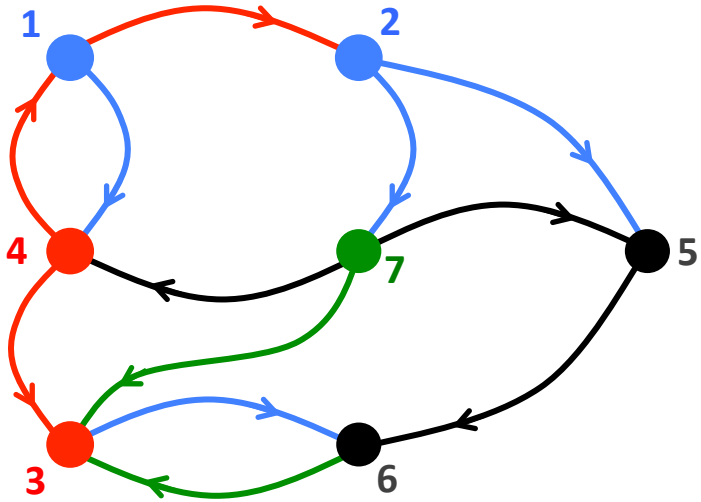
ALGORITHM:

1. Find owners of the current frontier's adjacency [computation]
2. Exchange adjacencies via all-to-all. **[communication]**
3. Update distances/parents for unvisited vertices. [computation]

2D parallel BFS algorithm

[B., Madduri, 2011]





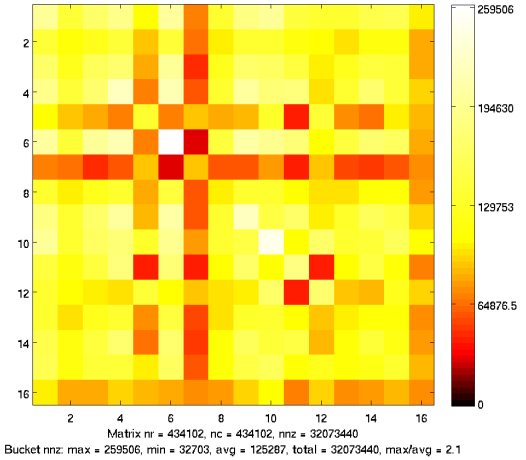
ALGORITHM:

1. Gather vertices in *processor column* **[communication]**
2. Find owners of the current frontier's adjacency [computation]
3. Exchange adjacencies in *processor row* **[communication]**
4. Update distances/parents for unvisited vertices. [computation]

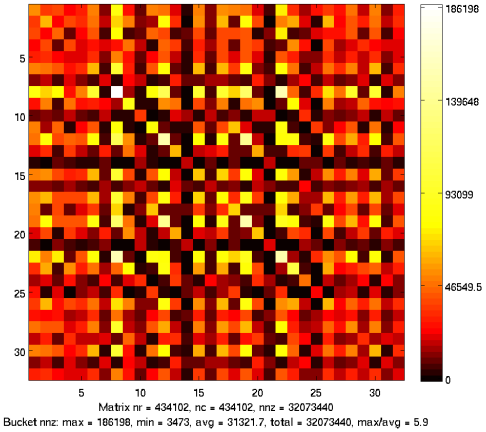
Orderings for the CoPapersCiteseer graph

[B, Madduri. Graph Partitioning for Scalable Distributed Graph Computations]

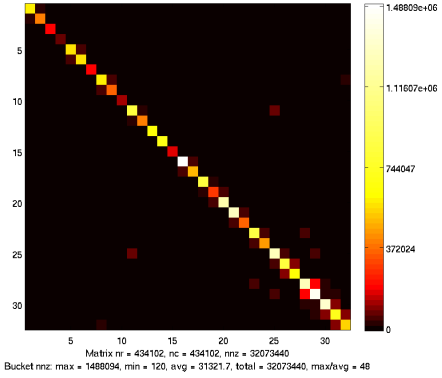
Natural



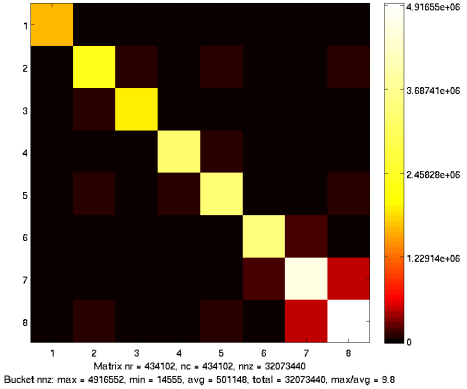
Random



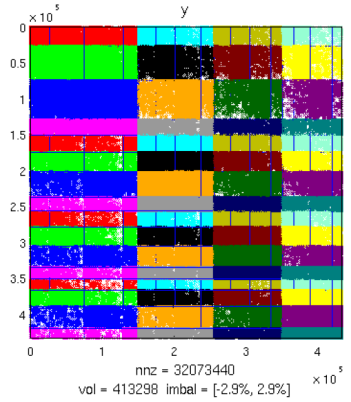
Metis



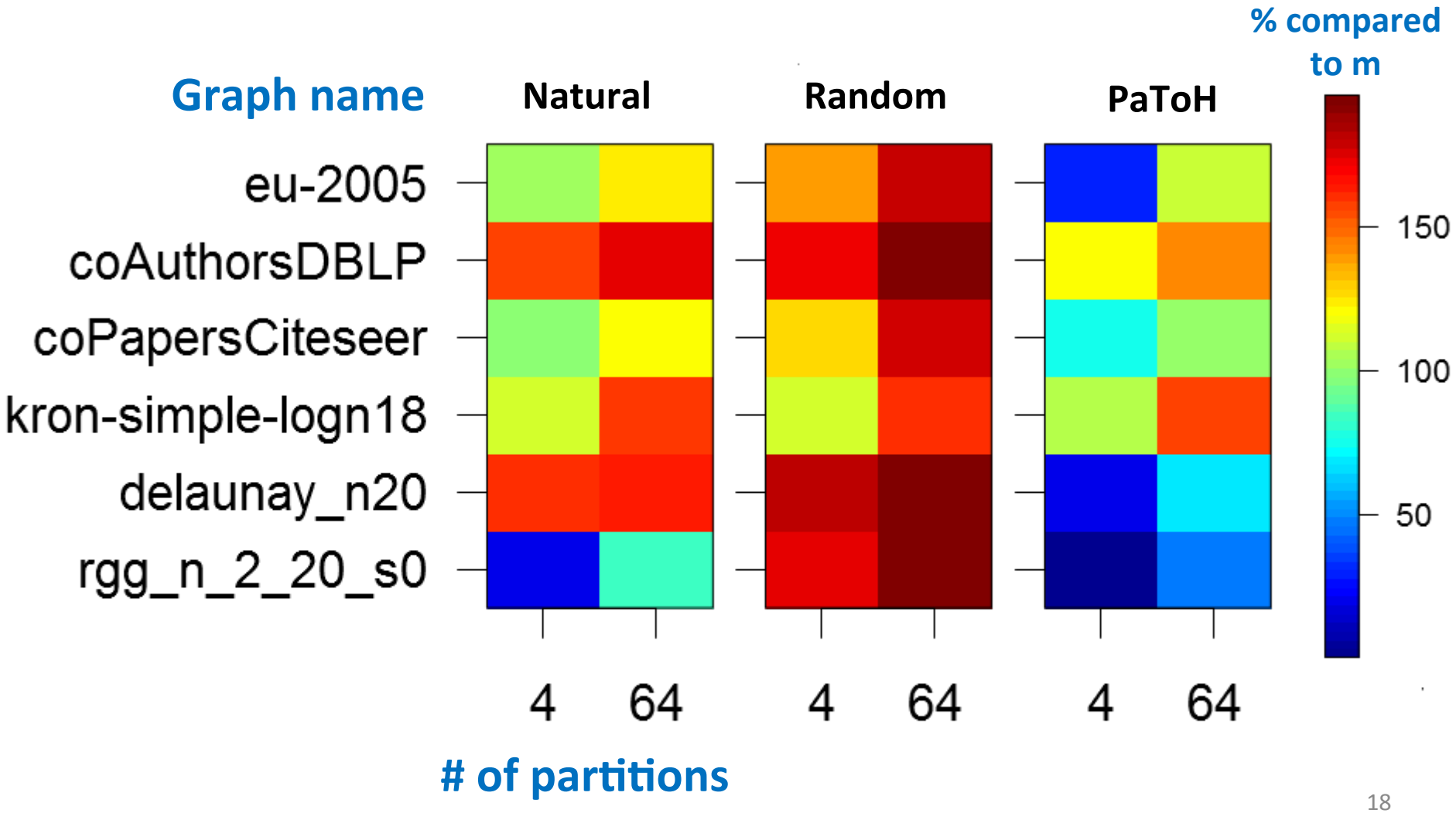
PaToH



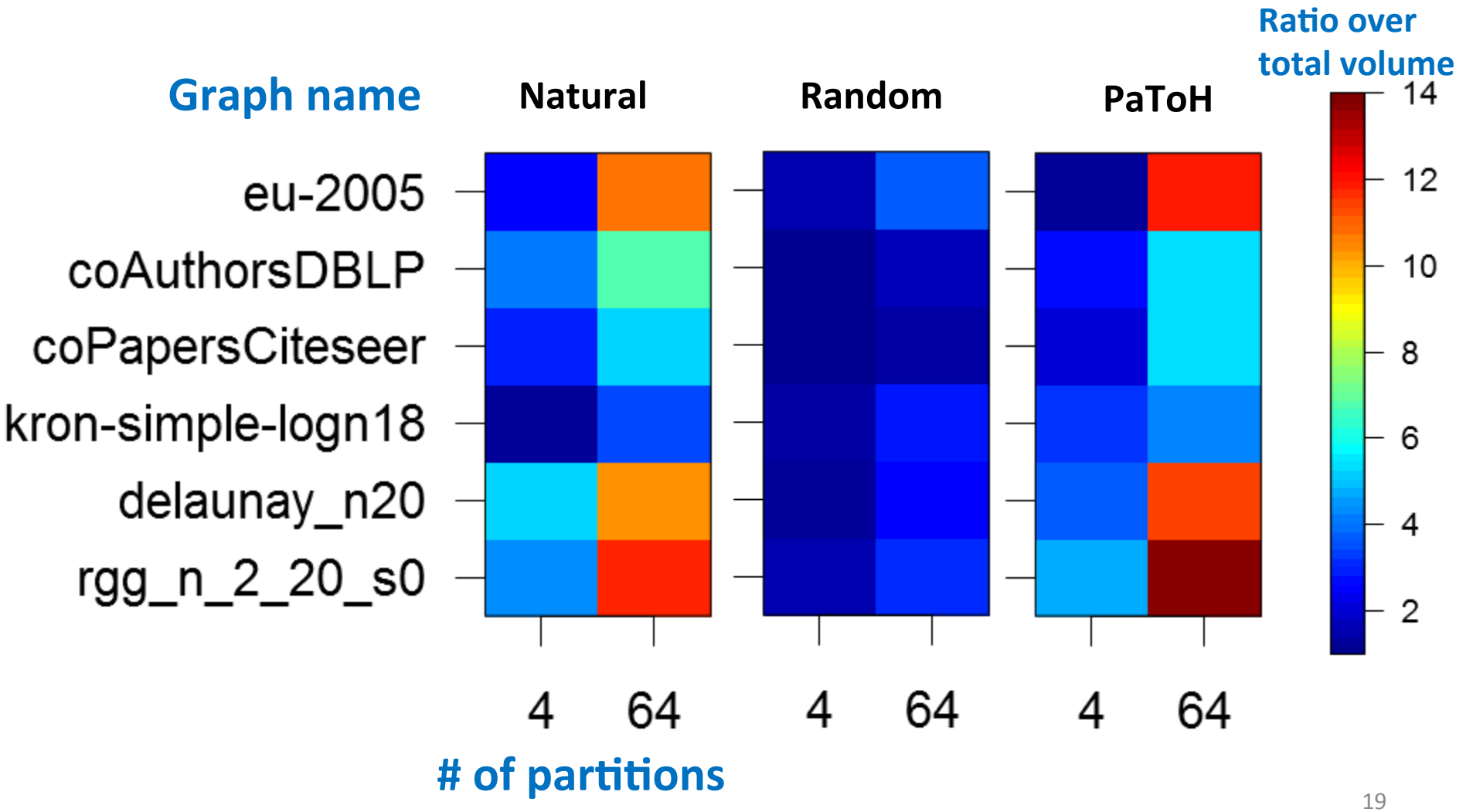
PaToH checkerboard



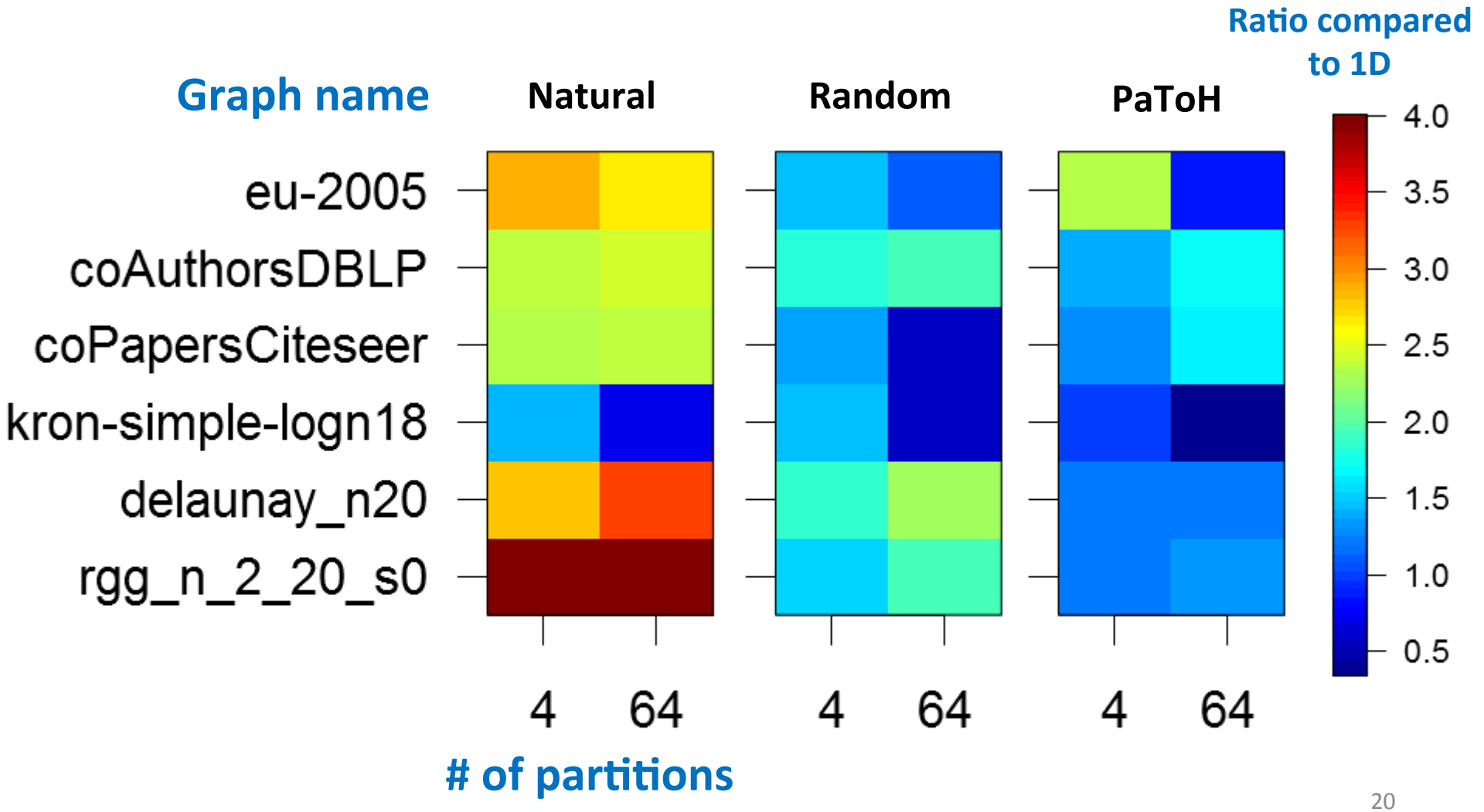
BFS All-to-all phase total communication volume normalized to # of edges (m)



Ratio of max. communication volume across iterations to average communication volume



Reduction in total All-to-all communication volume with 2D partitioning



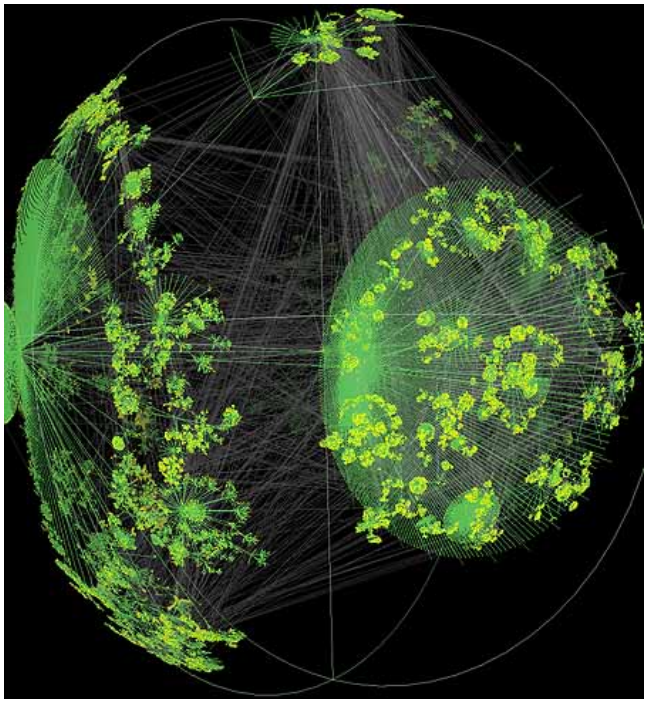
Outline

- The bottleneck: Communication
- The problem with graph partitioners
- Architectural evolution
- Data diversity: Graph characteristics
- Algorithmic evolution: Beat the worst case
- Functionality evolution: Most important kernels
- Peripherals: Database/visualization integration

Large graphs are everywhere

Internet structure

Social interactions

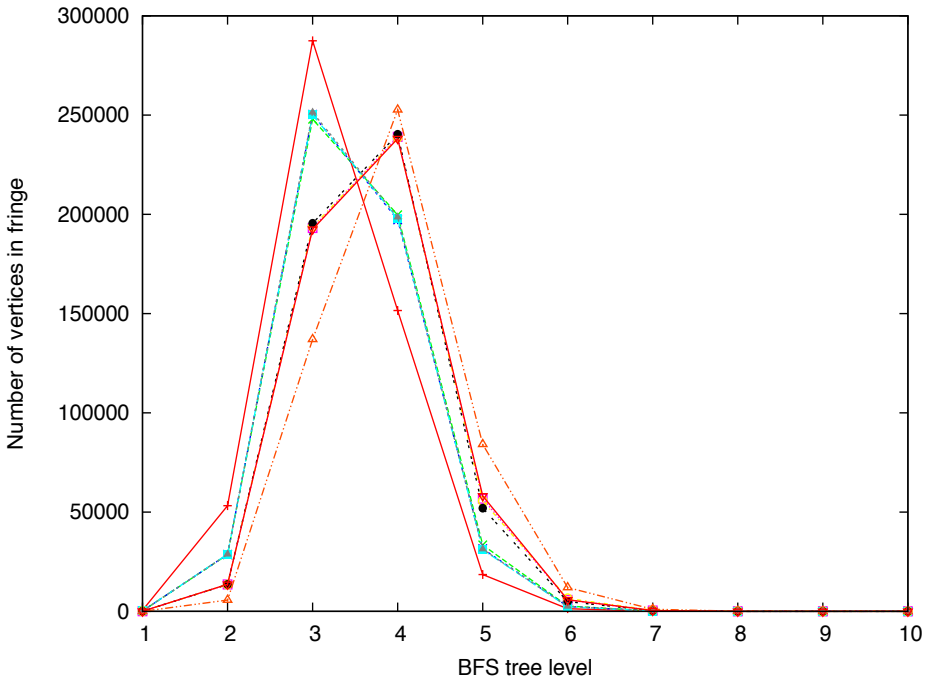


WWW snapshot, courtesy Y. Hyun

Scientific datasets: biological, chemical, cosmological, ecological, …

Yeast protein interaction network, courtesy H. Jeong

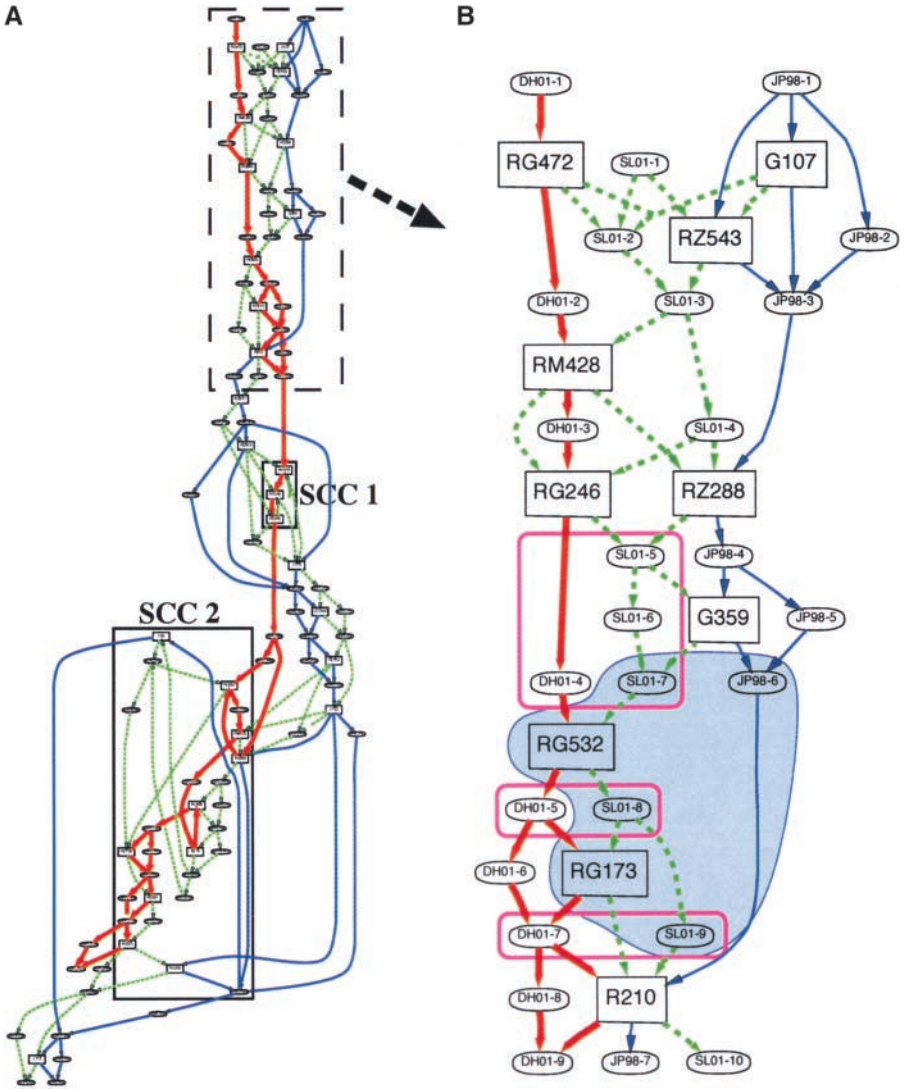
But they are NOT the same



Low diameter R-MAT graph

vs.

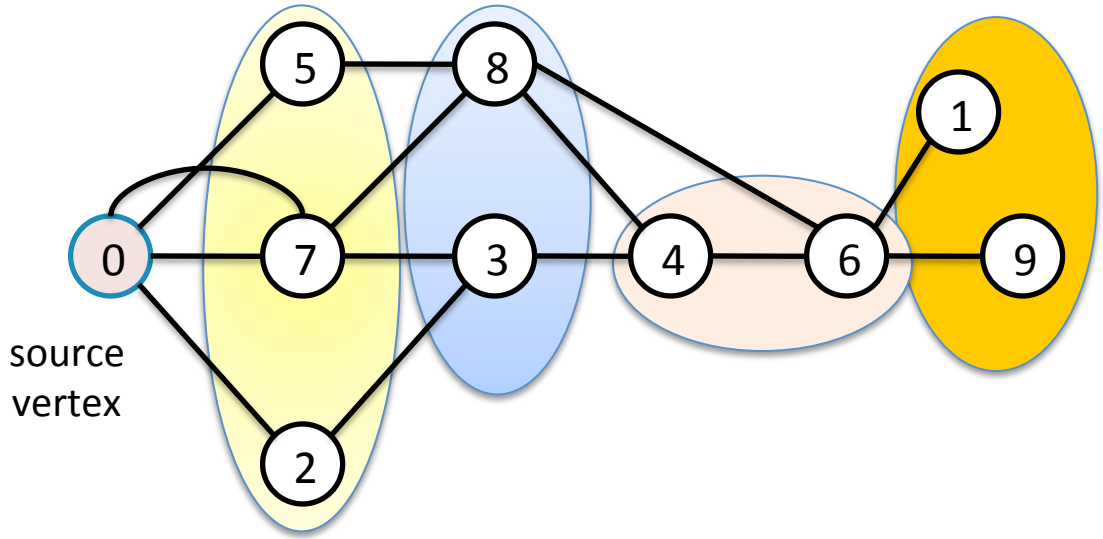
Long skinny genome graph



Gene linkage map, courtesy Yan et al.

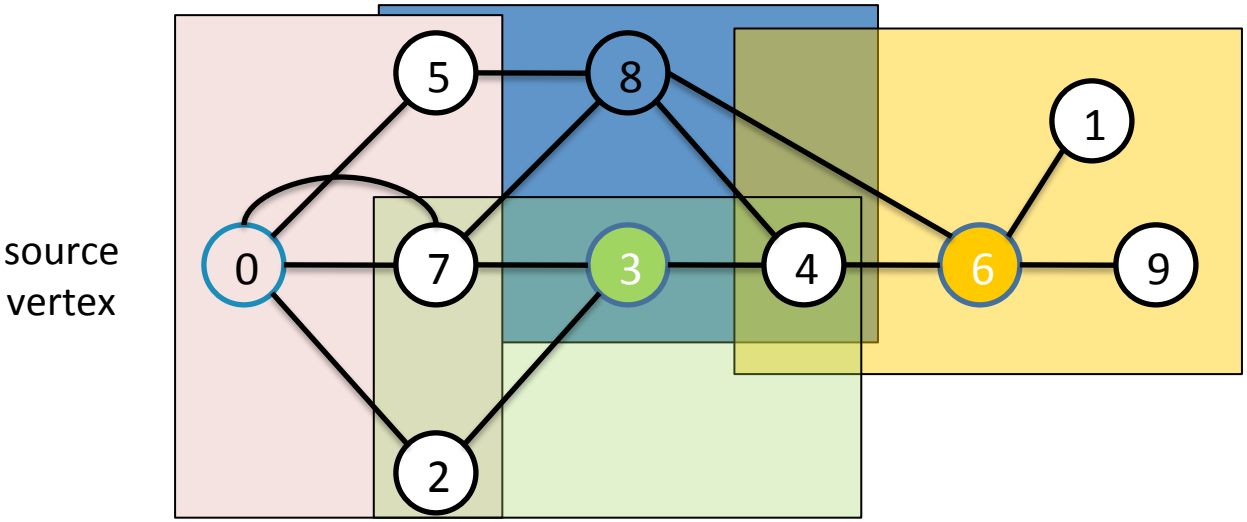
Parallel BFS strategies

1. Expand current frontier (**level-synchronous** approach, suited for **low diameter** graphs)



- O(D) parallel steps
- Adjacencies of all vertices in current frontier are visited in parallel

2. Stitch multiple concurrent traversals (Ullman-Yannakakis, for **high-diameter** graphs)



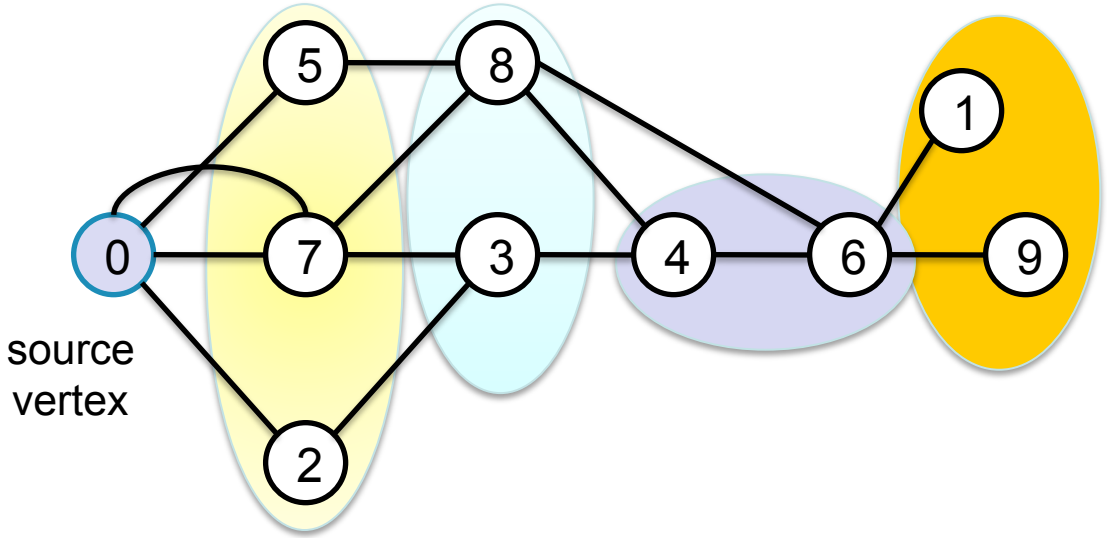
- path-limited searches from "super vertices"
- APSP between "super vertices"

Outline

- The bottleneck: Communication
- The problem with graph partitioners
- Architectural evolution
- Data diversity: Graph characteristics
- Algorithmic evolution: Beat the worst case
- Functionality evolution: Most important kernels
- Peripherals: Database/visualization integration

Bottom-up BFS

[Beamer, Asanović, Patterson, 2011]



BFS does not have to be O(m) all the time !

Step	Frontier Size	Fraction of Runtime	Edge Examinations	Failed Attempts	Fraction Failed
0	1	0.00002	242	0	0
1	242	0.01836	5,055,487	2,031,553	0.402
2	3,023,934	0.63358	2,902,729,050	2,847,737,876	0.981
3	54,991,174	0.32917	1,309,552,404	1,304,547,038	0.996
4	5,005,366	0.01755	5,870,543	5,855,182	0.997
5	15,361	0.00133	15,406	15,368	0.997
6	38	0.00001	38	38	1.0
Total	63,036,116	1.0	4,223,223,170	4,160,187,055	0.985

Acknowledgments…

David Bader, Grey Ballard, Scott Beamer, Ceren Budak, James Demmel, Armando Fox, John Gilbert, Laura Grigori, Bruce Hendrickson, Shoaib Kamil, Jeremy Kepner, Adam Lugowski, Kamesh Madduri, Lenny Oliker, Steve Reinhardt, Viral Shah, Oded Schwartz, Edgar Solomonik, Sam Williams