ON NETWORK LOCALITY IN MPI-BASED HPC APPLICATIONS

Felix Zahn, Holger Fröning

49th International Conference on Parallel Processing - ICPP

August, 20th 2020, Edmonton, AB, Canada

MOTIVATION

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

"Locality is efficiency,
Efficiency is power,
Power is performance,
Performance is king"

- Bill Dally*

*according to Wikipedia / cited from

Johnson, Matt (2011). An Analysis of Linux Scalability to Many Cores. p. 4

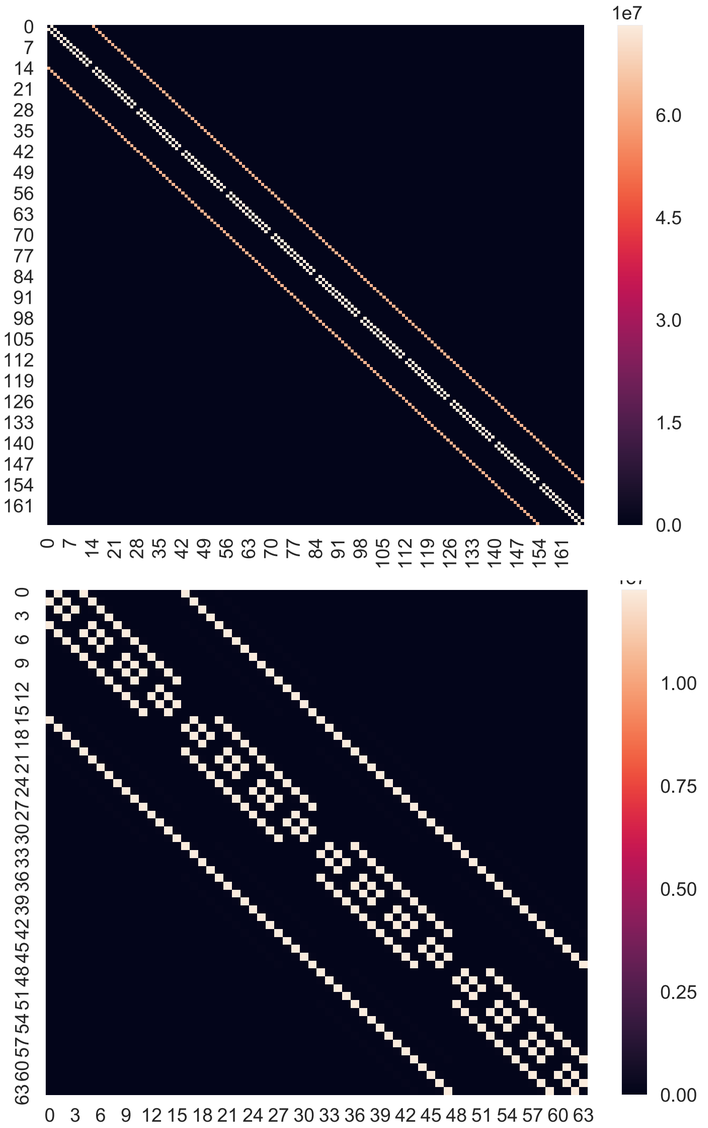
MPI TRAFFIC PATTERNS

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Currently:

- Message/injection rate, message size distribution, idle times, …
- Heat maps
- Good “human readability”

=> How to compare workloads objectively?



NEW METRICS

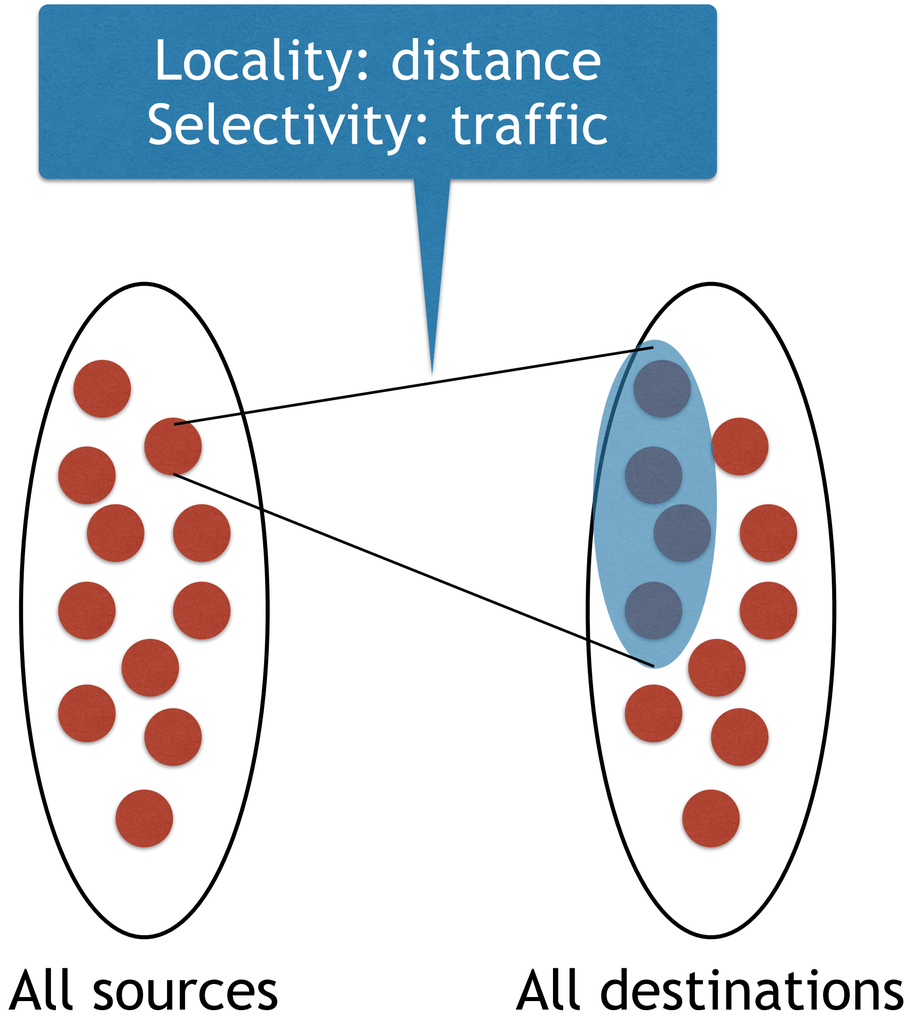
UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Application level:

- Directly derived from MPI traces
- No particular network information
- **Metric: Locality**
- **Metric: Selectivity**

Network level:

- Model for torus, fat-tree, and dragonfly
- Studies include multi-core analyses, network locality, and network utilization



APPLICATIONS

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

DoE exascale mini-applications

- Repository with dumpi traces
- maintained by Sandia National Laboratories

Limitations

- Missing header (no information about MPI DDTs)
- Custom communicators (MPI_Cart_create, MPI_Cart_sub)

Application	Description	Comm. Pattern
MOCFE (CESAR)	Neutronics code.	Nearest neighbor
NEKBONE (CESAR)	Fluid Dynamics code	Nearest neighbor
CNS (EXACT)	compressed Navier-Stokes equation	Nearest neighbor
MultiGrid (EXACT)	MultiGrid solver (BoxLib)	Nearest neighbor
LULESH (EXMATEX)	Hydrodynamic simulation	Nearest neighbor
CMC (EXMATEX)	Classic Monte Carlo	Nearest neighbor
AMG (DF)	Algebraic MultiGrid Solver	Nearest neighbor
AMR Boxlib (DF)	Adaptive mesh refinement	Irregular (sparse)
MiniAMR (DF)	Adaptive mesh refinement	Irregular (sparse)
BigFFT (DF)	large 3D FFT	Many-to-Many
Crystal Router (DF)	MPI many-to-many code	staged All2All
Fill Boundary (DF)	Halo update (BoxLib)	Nearest neighbor
MultiGrid (DF)	MultiGrid solver (BoxLib)	Nearest neighbor
MiniDFT (DF)	VASP electron structure calculation	Many2Many
MiniFE (DF)	Finite element solver	staged All2All
SNAP (DF)	Neural particle transport	Nearest neighbor
PARTISN (DF)	Neural particle transport	Nearest neighbor

https://portal.nersc.gov/project/CAL/doe-miniapps.htm

B. Klenk et al., "Relaxations for High-Performance Message Passing on Massively Parallel SIMT Processors," *2017 IPDPS*

RANK LOCALITY

"Distance" between rank pairs

$$dist = |rank_{source} - rank_{dest}|$$

$$locality = \frac{1}{dist}$$

90% threshold

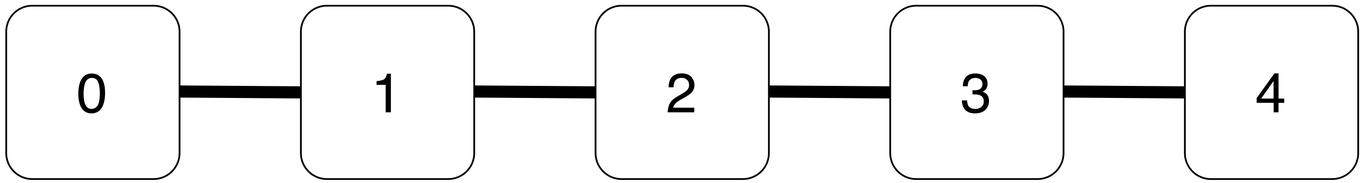
Hints communication pattern

Rank_0 -> Rank_2:

distance=2, locality=50%

Rank_4 -> Rank_1:

distance=3, locality=33%

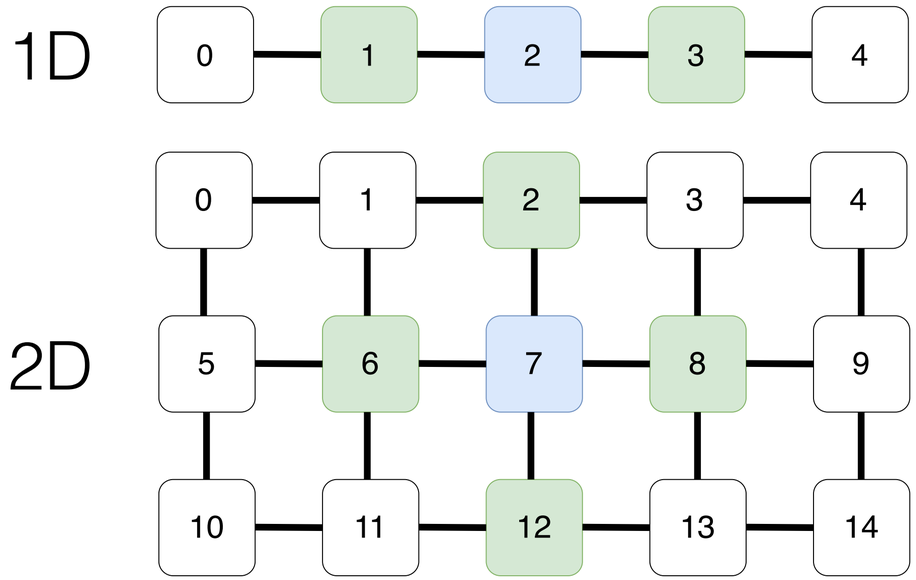


RANK LOCALITY

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Example nearest neighbor:

- Max. distance: 2 => locality > 50%
- Dimensionality



Workload	Ranks	Rank Locality 1D	2D	3D
AMG	216	3%	17%	100%
	1728	1%	8%	100%
Boxlib CNS large	64	3%	13%	21%
	256	1%	8%	13%
	1024	0%	3%	7%
EXMATEX LULESH	64	6%	24%	100%
	512	2%	6%	100%
MultiGrid_C	125	2%	6%	17%
	1000	0%	3%	9%
PARTISN	168	7%	100%	22%

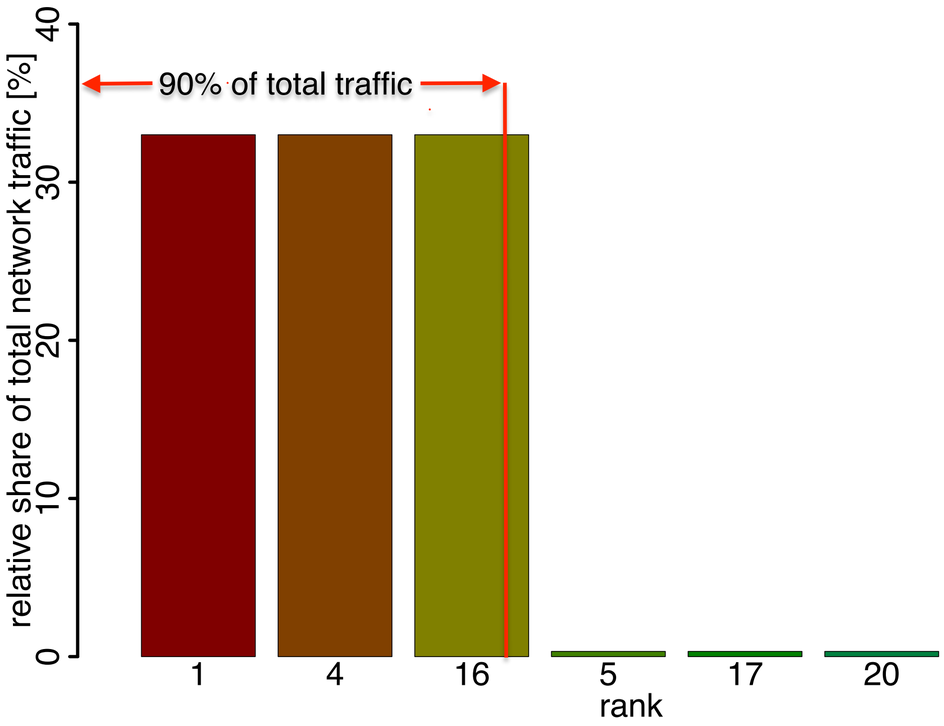
SELECTIVITY

How many communication partner does one rank send 90% of its traffic volume to

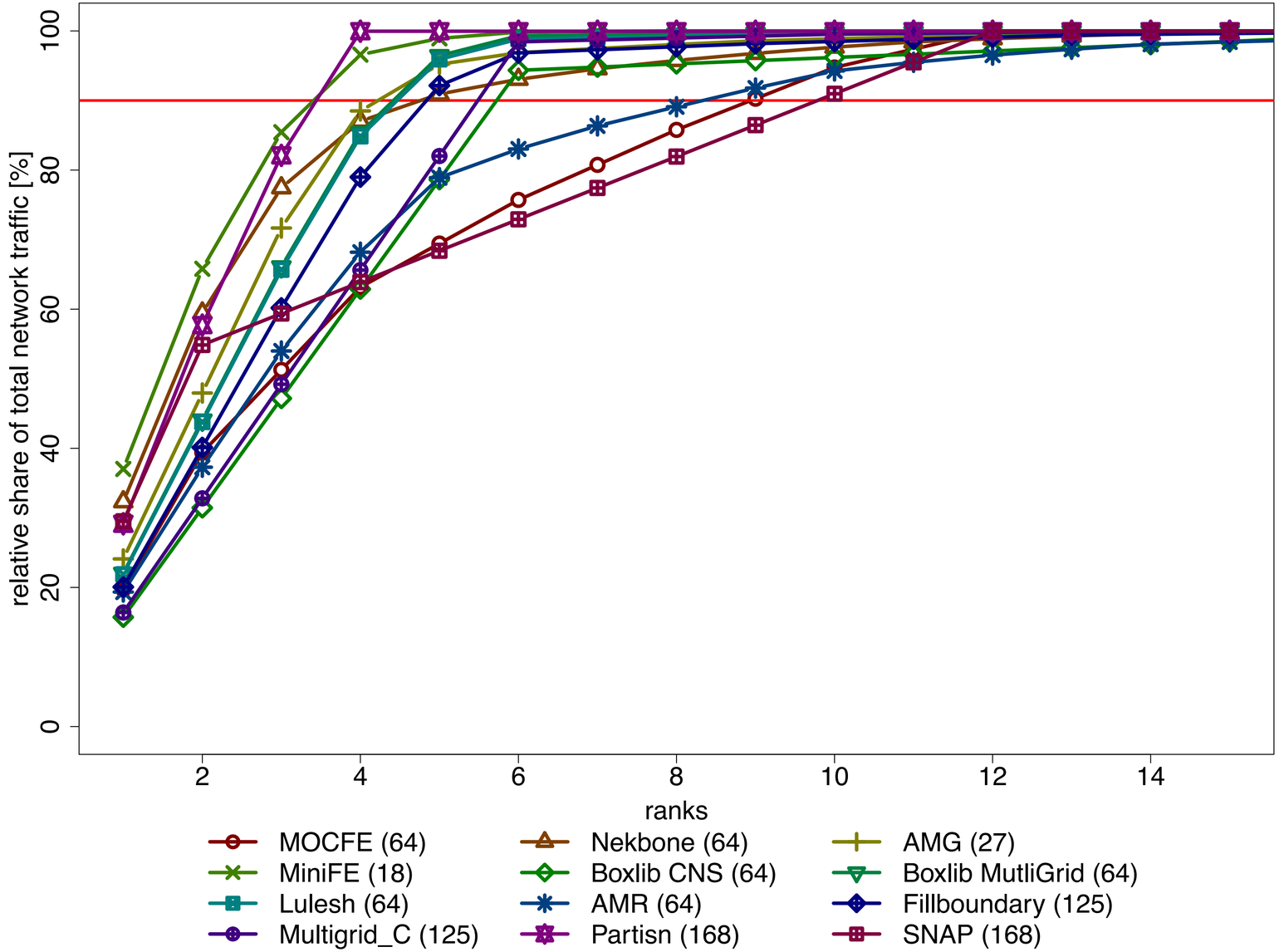
- Only point-to-point
- Independent from distances

Peers:

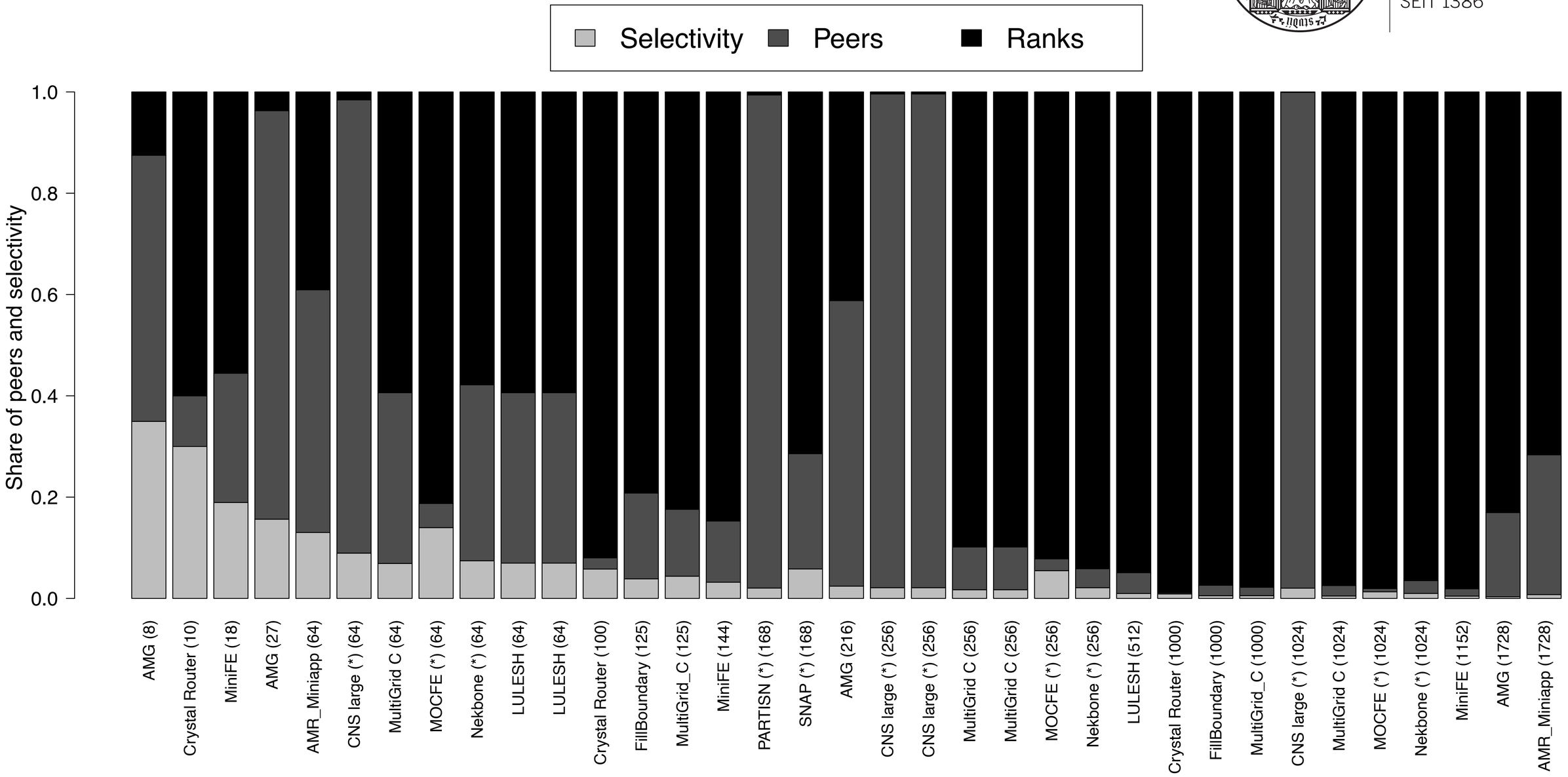
- Total number of communication partners
- For comparison (difference equals 10% of total traffic volume)



SELECTIVITY



SELECTIVITY



NETWORK LAYER

Model to provide static analyses

- No simulation needed, non-temporal character
- Including: routing, # links, BW, packets

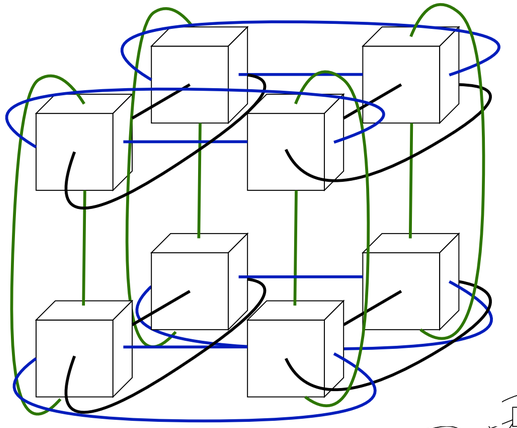
Topologies

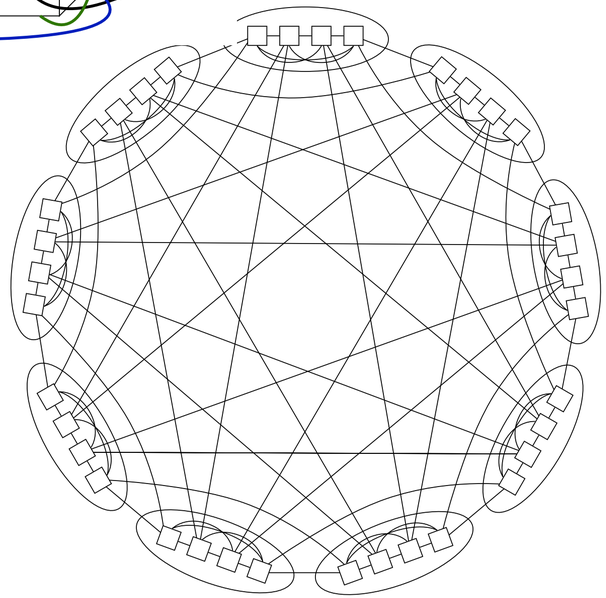
- 3D Torus
- Fat Tree
- Dragonfly

Shortest-path routing

- Deterministic
- No traffic flows / emphasis of topology characteristics

Incremental mapping

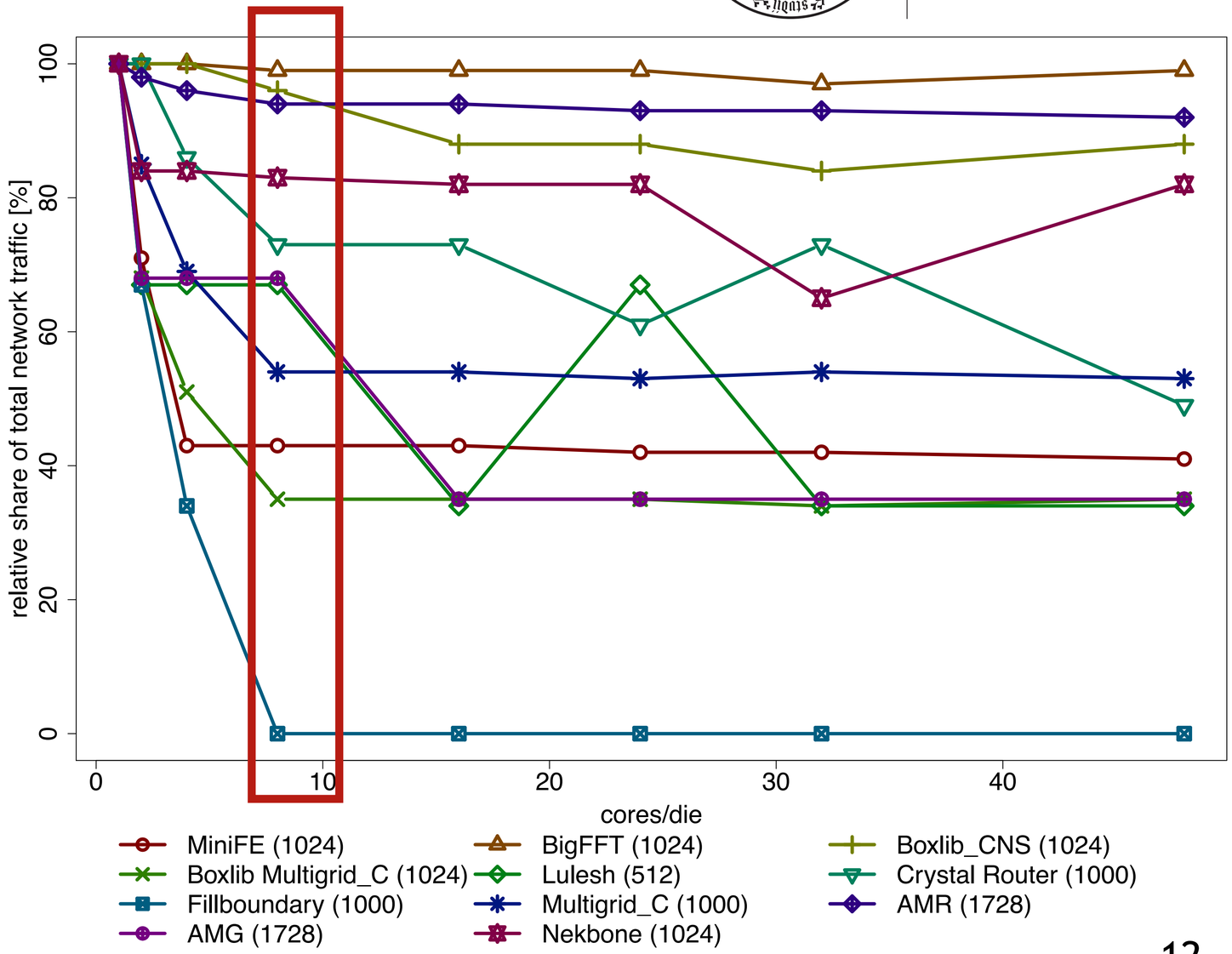




MULTI-CORE EFFECTS

How does network traffic change with the number of cores/socket?

- One rank/core
- Incremental mapping
- Includes p2p & collectives



HOPS & UTILIZATION

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Topological locality:

- Routing information required

$$packethops = \sum_{p \in packets} \#hops(p)$$

$$\overline{\text{hops}} = \frac{packethops}{\#packets}$$

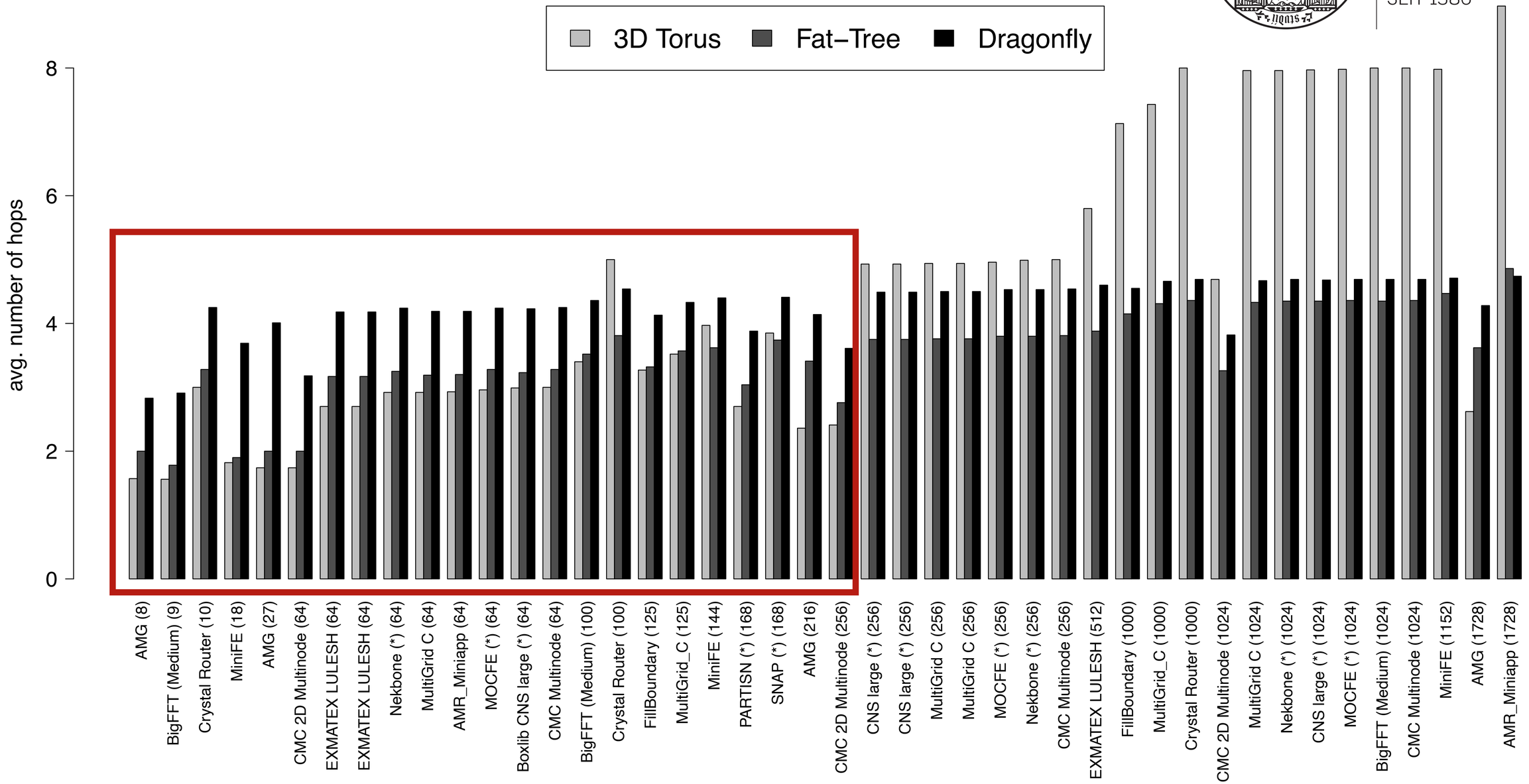
Network utilization:

- Depending on distances and number of links

$$datavolume = \sum_{p \in packets} \#hops(p) \cdot size(p)$$

$$utilization = \frac{datavolume}{BW \cdot t_{execution} \cdot \#links}$$

HOPS & UTILIZATION



UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

HOPS & UTILIZATION

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Workload	Ranks	Network utilization [%] 3D Torus	[%] fat-tree	[%] Dragonfly
AMG	8	0.0052	0.0303	0.0116
	27	0.0012	0.0034	0.0034
	216	0.0008	0.0032	0.0021
	1728	0.0001	0.0004	0.0002
AMR_Miniapp	64	0.0034	0.0058	0.0048
	1728	0.0278	0.0229	0.0119
BigFFT (Medium)	9	0.6721	3.0725	1.2943
	100	7.4849	10.5544	7.6985
	1024	47.2317	38.4346	22.1491
Boxlib CNS large (*)	64	0.0002	0.0003	0.0003
	256	0.0004	0.0005	0.0004
	256	0.0005	0.0006	0.0004
	1024	0.0012	0.0010	0.0006
Boxlib MultiGrid C	64	0.0011	0.0020	0.0017
	256	0.0035	0.0045	0.0032
	256	0.0036	0.0046	0.0033
	1024	0.0106	0.0092	0.0054
CESAR MOCFE (*)	64	0.0498	0.0769	0.0605
	256	0.1216	0.1368	0.0895
	1024	0.4495	0.3656	0.2108
CESAR Nekbone (*)	64	0.0027	0.0090	0.0081
	256	0.3447	0.3882	0.2541
	1024	0.0029	0.0057	0.0035

Workload	Ranks	Network utilization [%] 3D Torus	[%] fat-tree	[%] Dragonfly
Crystal Router	10	0.0469	0.1938	0.0882
	100	0.0408	0.0637	0.0490
	1000	0.1475	0.1531	0.0959
EXMATEX CMC 2D Multinode	64	2.0E-05	3.0E-05	2.4E-05
	256	0.0001	0.0001	0.0001
	1024	0.0008	0.0007	0.0004
EXMATEX LULESH	64	0.0004	0.0013	0.0011
	64	0.0004	0.0016	0.0013
	512	0.0005	0.0020	0.0012
FillBoundary	125	0.0319	0.0466	0.0351
	1000	0.0245	0.0248	0.0160
MiniFE	18	0.0008	0.0031	0.0015
	144	0.0017	0.0025	0.0017
	1152	0.0039	0.0037	0.0022
MultiGrid_C	125	0.0038	0.0056	0.0041
	1000	0.0013	0.0013	0.0008
PARTISN (*)	168	7.4E-08	1.6E-07	1.2E-07
SNAP (*)	168	4.2E-07	6.2E-07	4.0E-07

CONCLUSION

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Two new metrics to quantify communication patterns

- **Rank Locality:** deriving types of communication and dimensionality
- **Selectivity:** small subsets of ranks cause most of communication

  -> advanced mapping can reduce network traffic

Network level analyses

- Multi-core: traffic plateau at 8 cores/socket -> no further reduction of traffic volume
- $\overline{\text{hops}}$: 3D torus best fit for <256 ranks, fat-tree undertakes for larger configurations
- Overall low network utilization of less than 1% for 14/15 applications.

OUTLOOK

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Analyzing more traffic patterns for locality and selectivity

- Goal: Predict communication pattern from two values

Further quantitative description of MPI communication

- Substitute complex network simulation with accurate models
- Deeper understanding benefits network design

Mapping

- Identifying subset of heavily communicating traces to map them together
- Reduce of network load and save energy by provisioning the network appropriately

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

THANK YOU!

SPONSORS

Carl Zeiss Stiftung