Visual Analytics Techniques for Exploring the Design Space of Large-Scale High-Radix Networks

Kelvin Li, Kwan-Liu Ma,

Misbah Mubarak, Robert Ross,

Christopher Carothers

IEEE Cluster 2017
Honolulu, Hawaii
September 6, 2017

Background

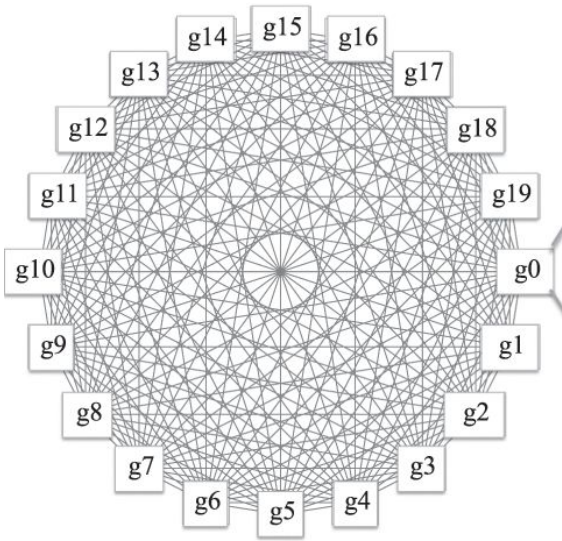
New Supercomputing Systems

- More compute nodes
- Increased network complexity

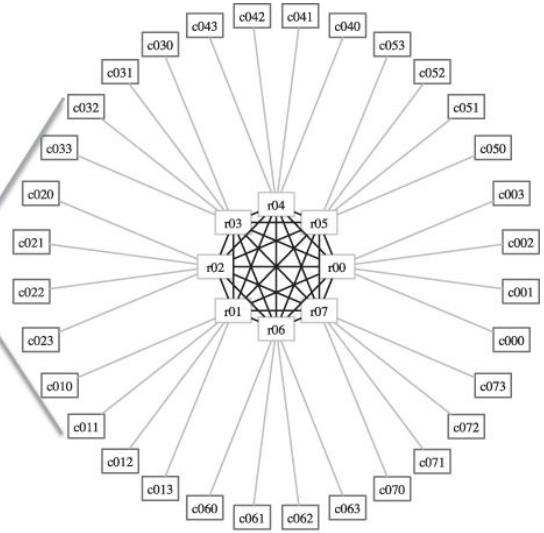
System	Topology	Nodes
Theta (ANL)	Dragonfly	> 3,000
Sierra (LLNL)	Fat Tree	> 3,400
Cori (NERSC)	Dragonfly	> 10,000
Trinity (LANL)	Dragonfly	> 19,000
Aurora (ANL)	Dragonfly	> 50,000

The Dragonfly Network Topology

- Kim et. al 2008
- A two-level direct connected topology
- A popular choices for building new supercomputers



Groups are fully connected (all-to-all) with global links.

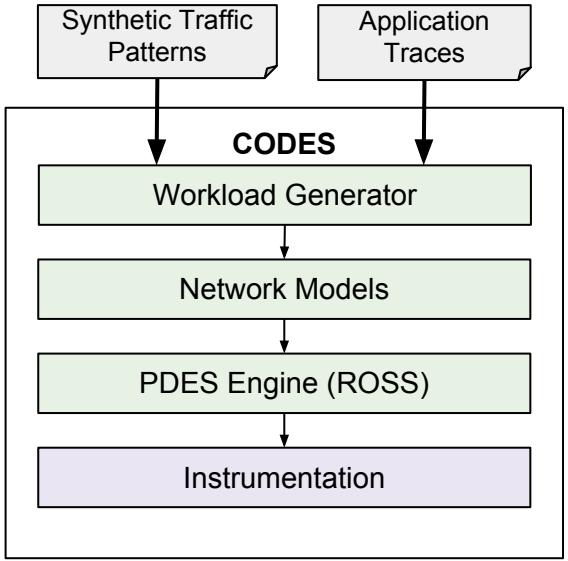


Routers within each group are fully connected via local links.

Motivation

- Parallel discrete-event simulation (PDES) provides a cost-effective way to evaluate designs of Dragonfly networks.
- Effective visualizations are useful for steering the analysis and exploration process.
- PDES + Visual Analytics provides a powerful toolkit for exploring the design space of large-scale HPC networks.

CODES Network Simulation Toolkit

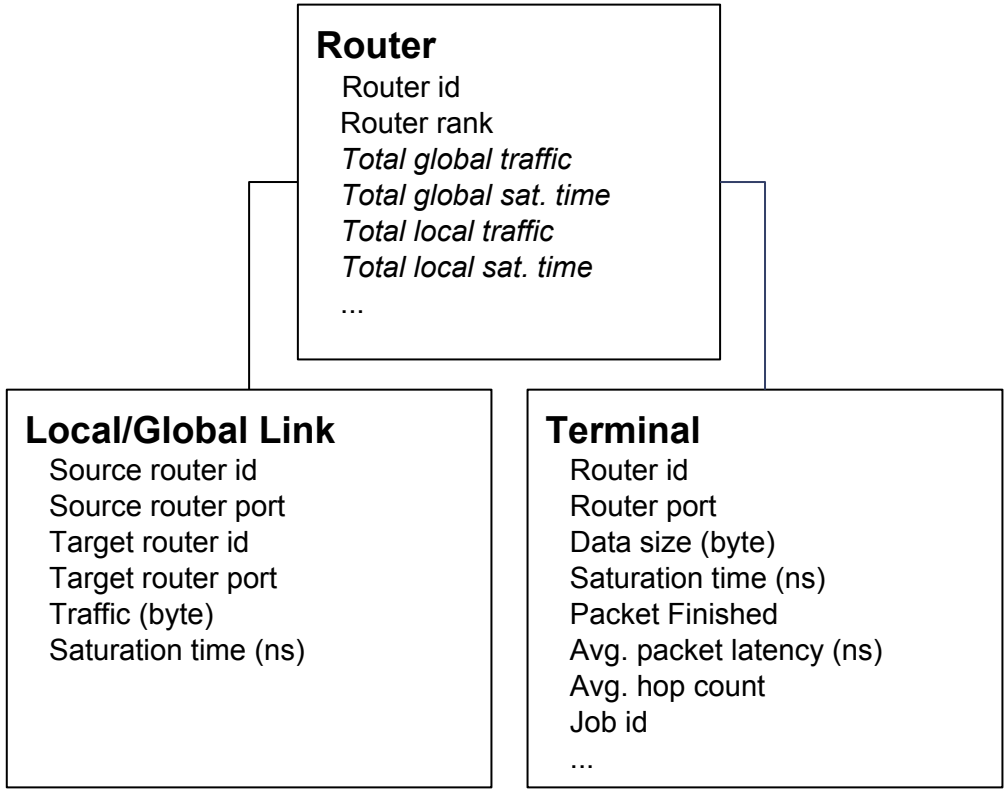


- High fidelity (packet-level detail) network simulation
- Synthetic and application trace workloads
- Different network configurations
  - Interconnect topologies
  - Routing strategies
  - Job placement policies

CODES website: **http://www.mcs.anl.gov/projects/codes/**

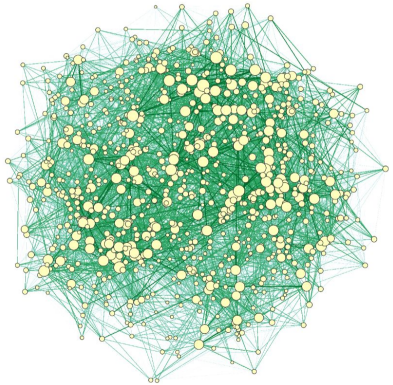
Data Collection in CODES

- Multiple entities: router, network links, and terminals
- Various performance metrics (e.g., data size, saturation time)
- Time varying

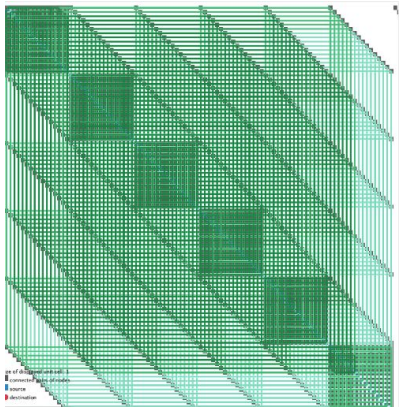


Challenges for Visualizing Large-Scale Networks

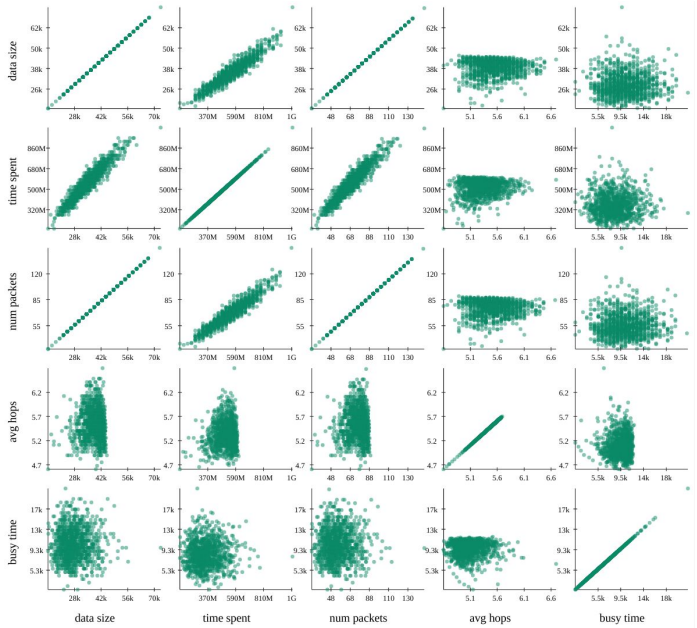
Simulation of a Dragonfly network with ~1,000 terminals.



Force-directed layout



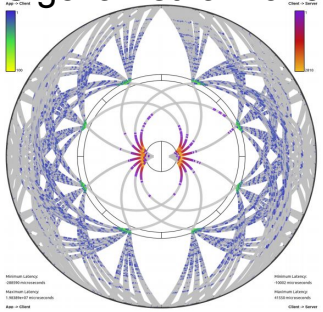
Adjacency matrix layout



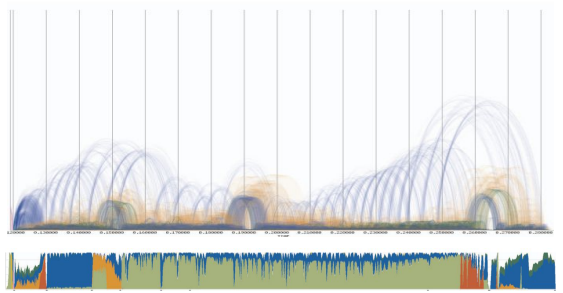
Visualization Tools for Network Analysis

- Designed for specific analysis tasks
- Fixed visualization layout
- Limited scalability

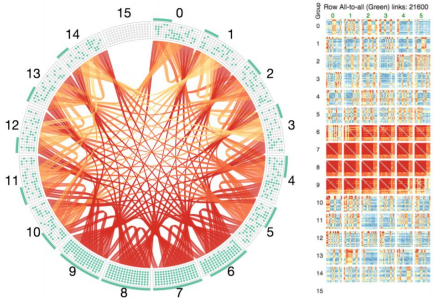
Sigovan et. al 2013



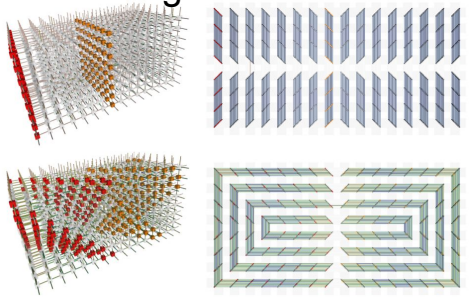
Muelder et. al 2009



Bhatele et. al 2015



Landge et. al 2012



Studying Dragonfly Network Performance

Exploring design choices

- Link arrangements
- Routing strategies
- Job placement policies

…

Analyzing network behaviors

- Communication pattern
- Network congestion
- Inter-job interference

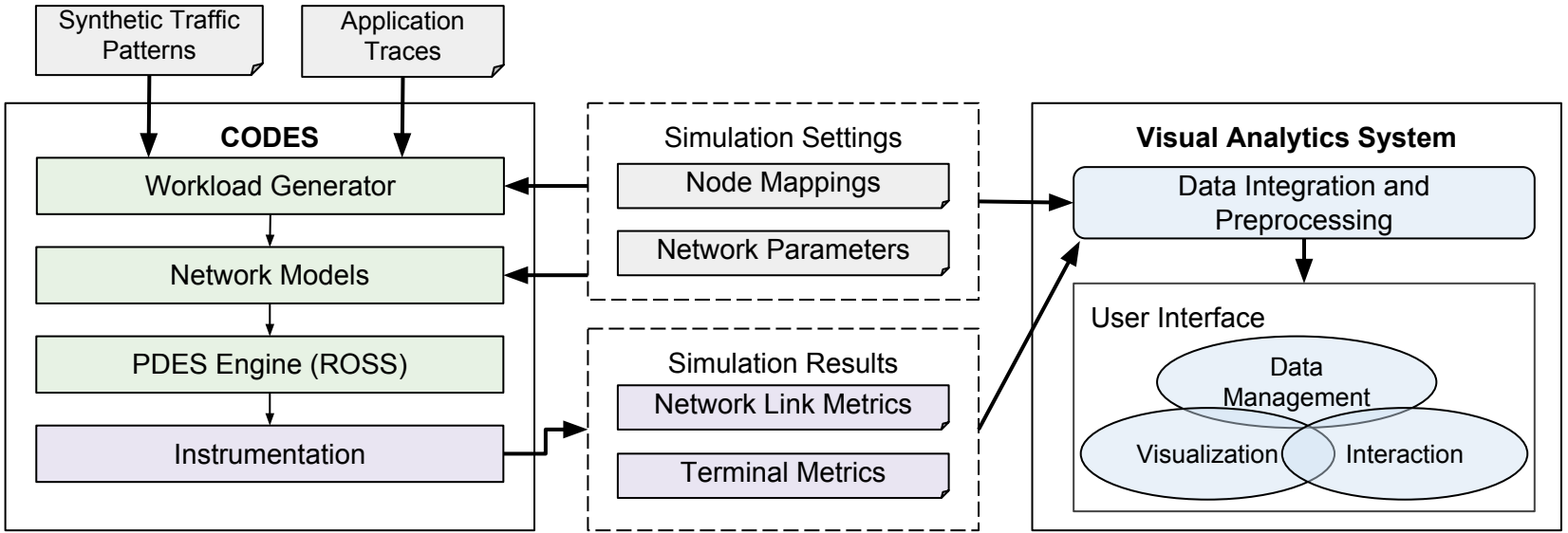
…

Objectives

To develop effective visual analytics solution for supporting design space exploration of large-scale Dragonfly networks.

- **Scalable visualization** for analyzing large-scale networks
- **Flexible exploration** for studying various network behaviors

Visual Analytics System for Supporting CODES

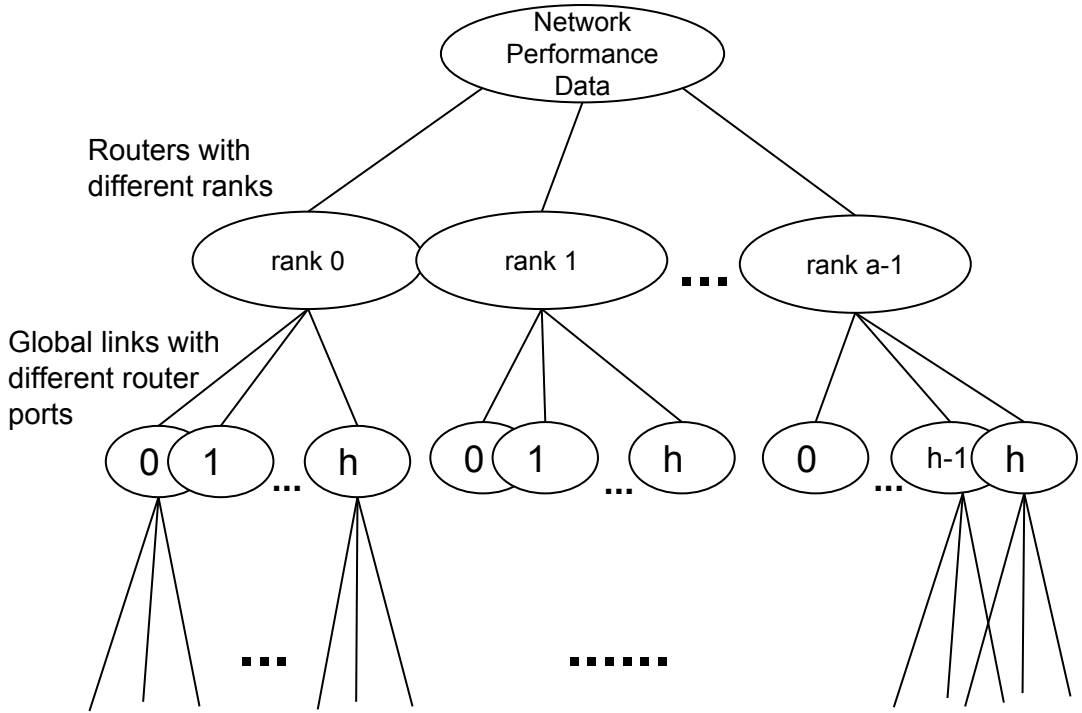


Visual Analytics Techniques

- Hierarchical Data Aggregation
  - Overview of the entire network
  - Different levels of details
- Radial visualization layout
  - Customized visual mappings
  - Correlation of performance

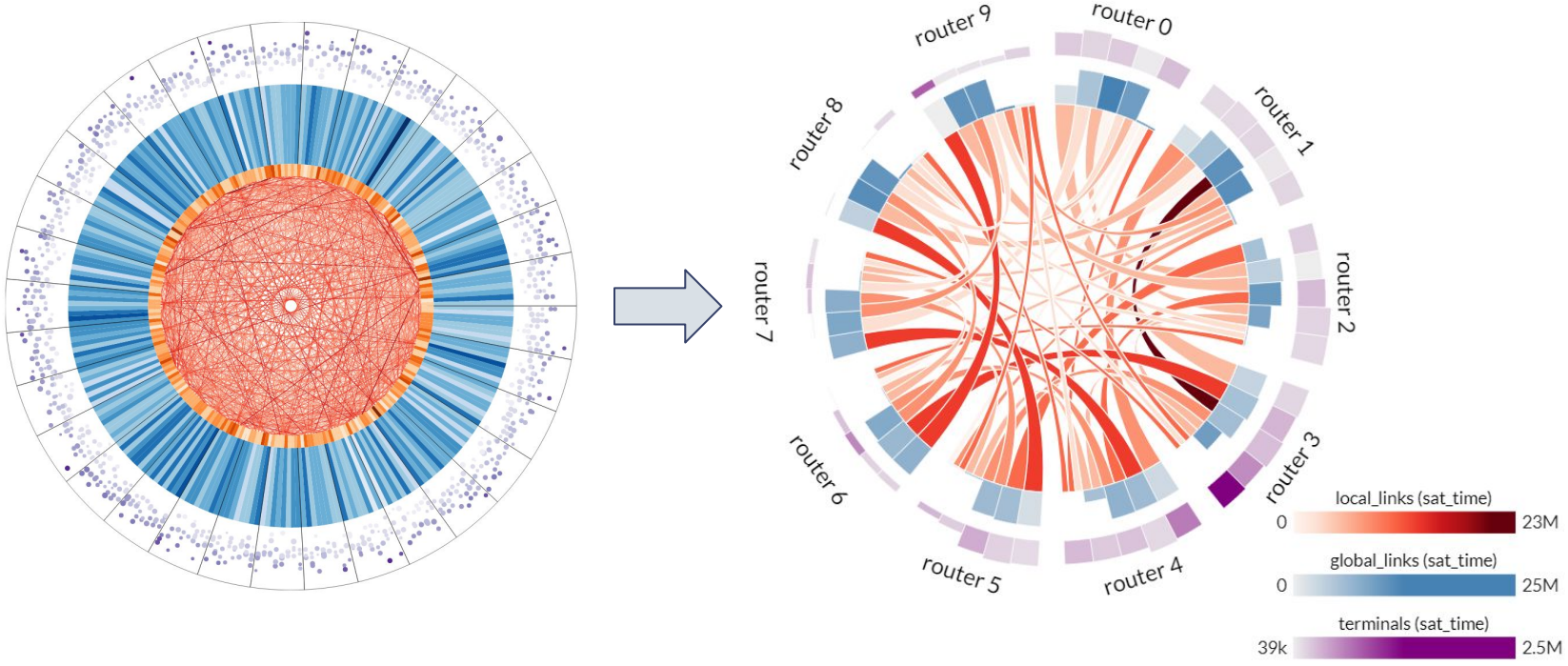
Hierarchical Aggregation for Visualization

- Construct an aggregate tree from data items for visualizing different levels of details
- Users choose the data attributes to be used for the aggregation at each tree level.

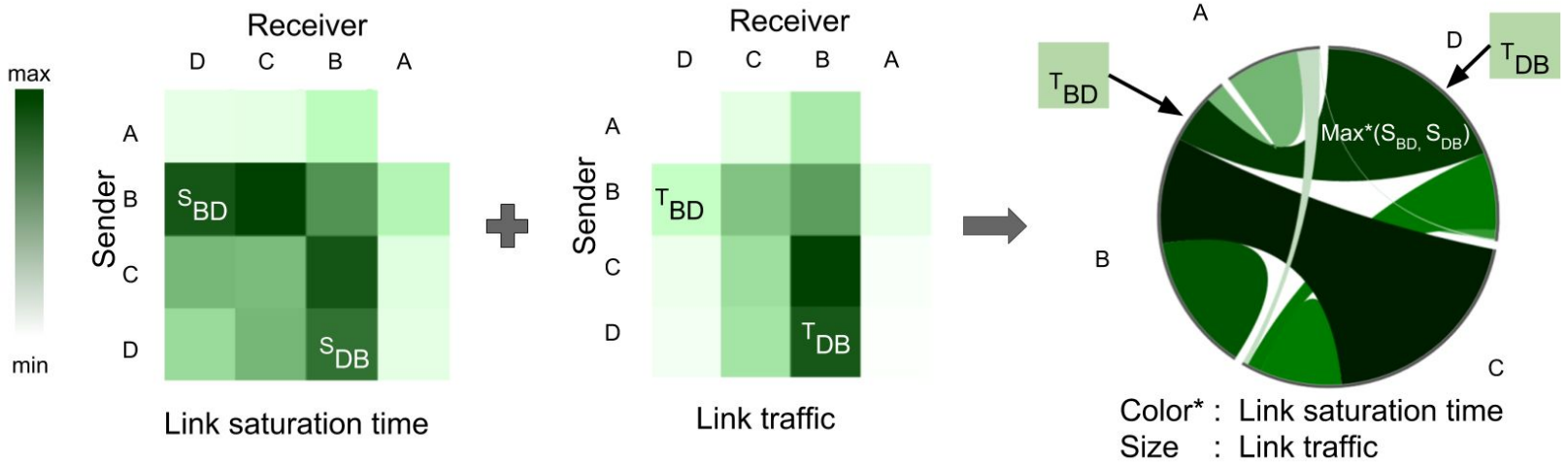


Example: Dragonfly network with a router per group and h global links per router

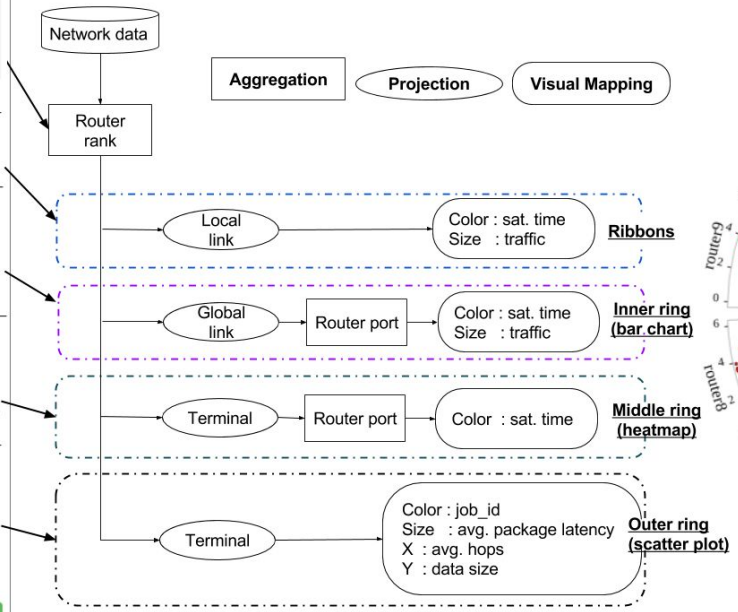
Hierarchical Radial Visualization

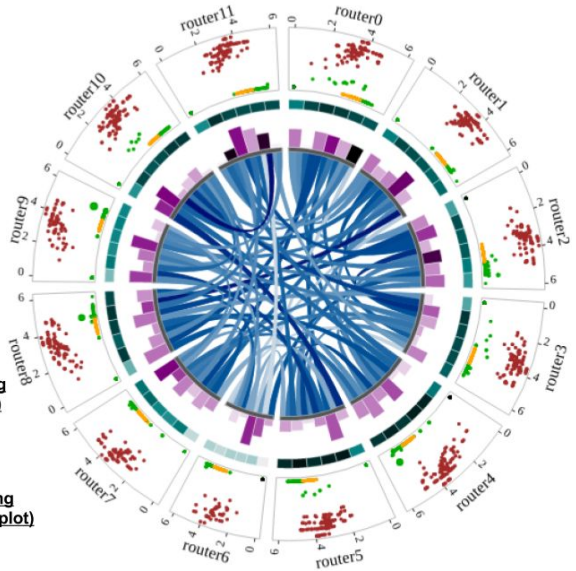


Visual Encoding for Network Links



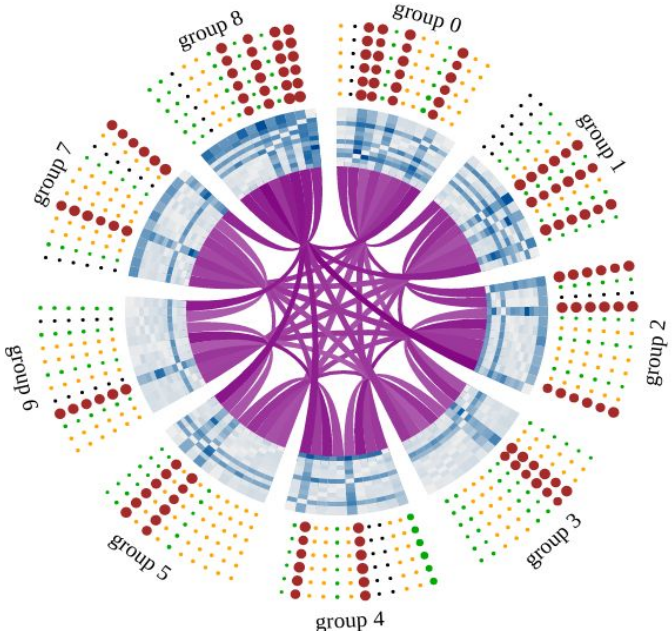
User Interface





Declarative Language

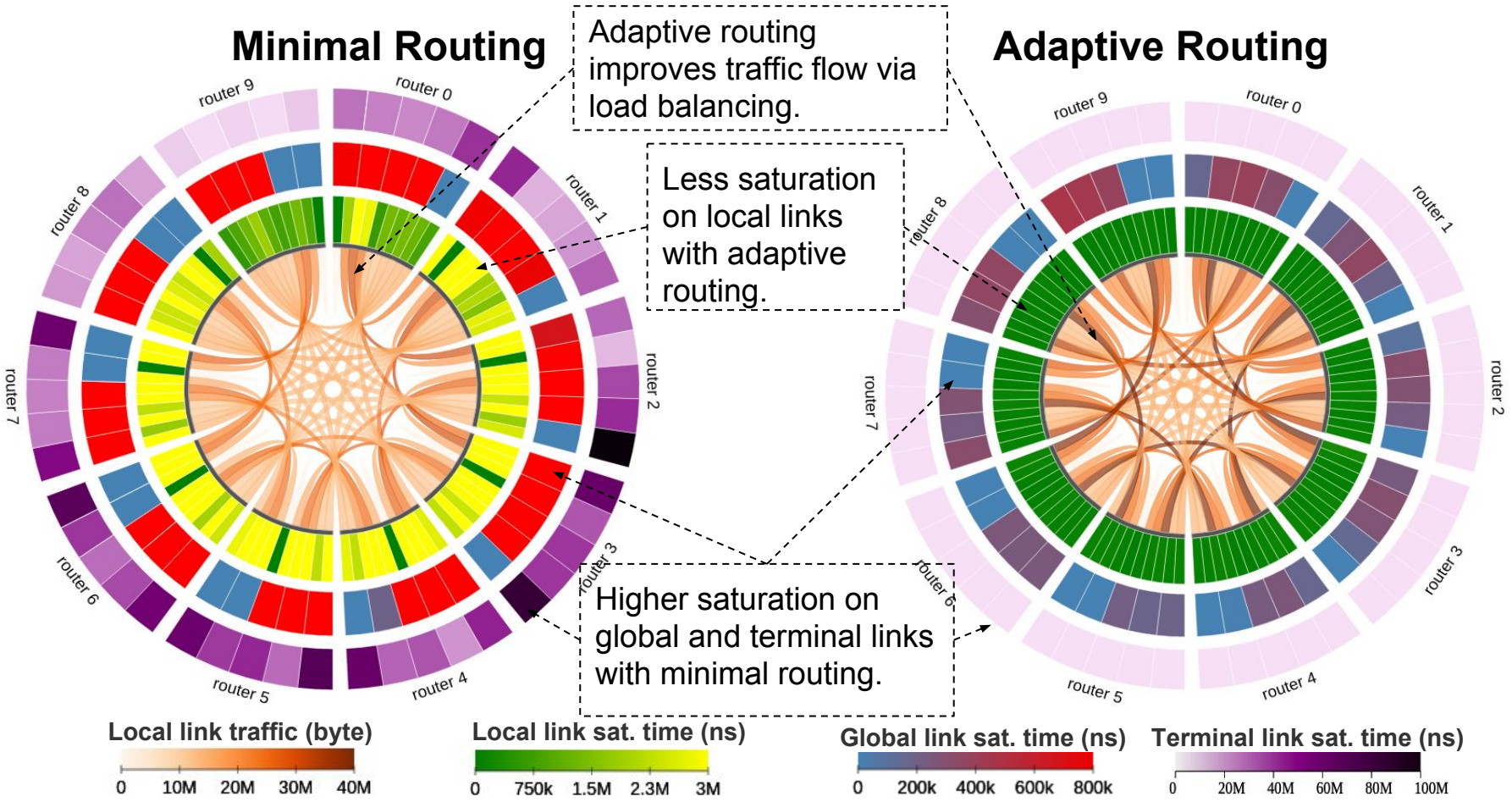
```
{
    filter: { group_id : [0, 8] },
    aggregate : "group_id",
    project : "global_links",
    vmap : { size : "traffic"},
    colors : ["white", "purple"]
},
{
    project : "local_links",
    vmap : {
        color : "traffic",
        x : "router_rank",
        y : "router_port"
    },
    colors : ["white", "steelblue"],
},
...
```



Demo

Comparing Routing Strategies

- Minimal vs. Adaptive
- Dragonfly network with 2,550 terminals
- Application: Algebraic Multigrid Solver (AMG) with a 3D nearest neighbor communication pattern


Minimal Routing
Adaptive Routing
Adaptive routing improves traffic flow via load balancing.
Less saturation on local links with adaptive routing.
Higher saturation on global and terminal links with minimal routing.
router 0
router 1
router 2
router 3
router 4
router 5
router 6
router 7
router 8
router 9
Local link traffic (byte)
0 10M 20M 30M 40M
Local link sat. time (ns)
0 750k 1.5M 2.3M 3M
Global link sat. time (ns)
0 200k 400k 600k 800k
Terminal link sat. time (ns)
0 20M 40M 60M 80M 100M

Exploring Inter-Job Interference

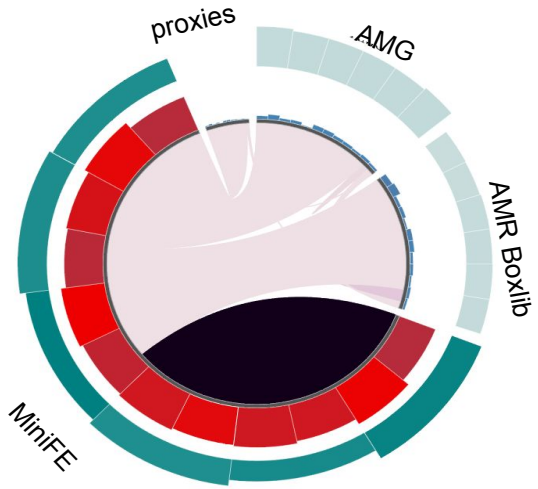
- 5,256 terminals in 73 groups, 12 routers per group
- Adaptive routing
- 3 applications running in parallel

Application	Ranks	Data	Comm. Pattern
AMG	1728	1.2GB	3D nearest neighbor
AMR Boxlib	1728	2.2GB	Irregular and sparse
MiniFE	1152	147GB	Many-to-many

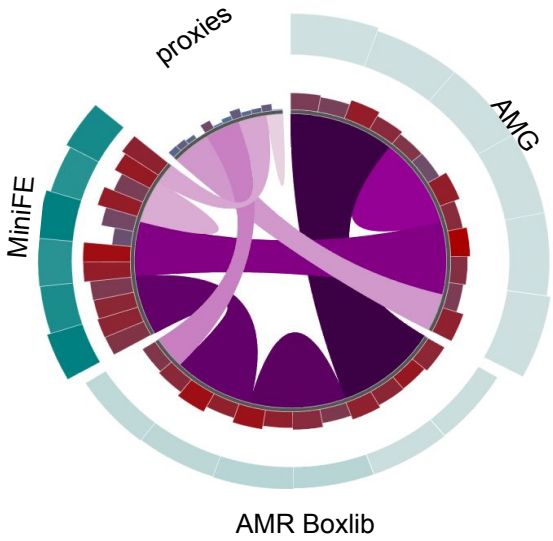
Comparing Job Placement Policies

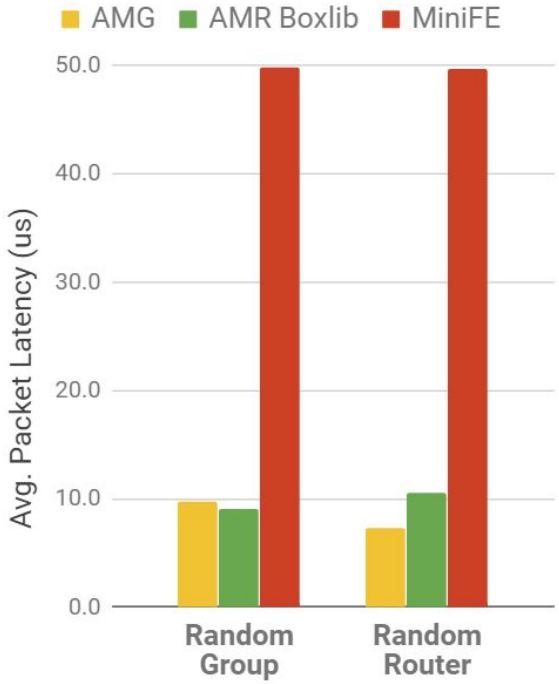
- Random Group
- Random Router
- Hybrid
  - Use random group placement for communication-intensive jobs
  - Use random router placement for jobs with less communication

Random Group



Random Router





Global link saturation

0 420us

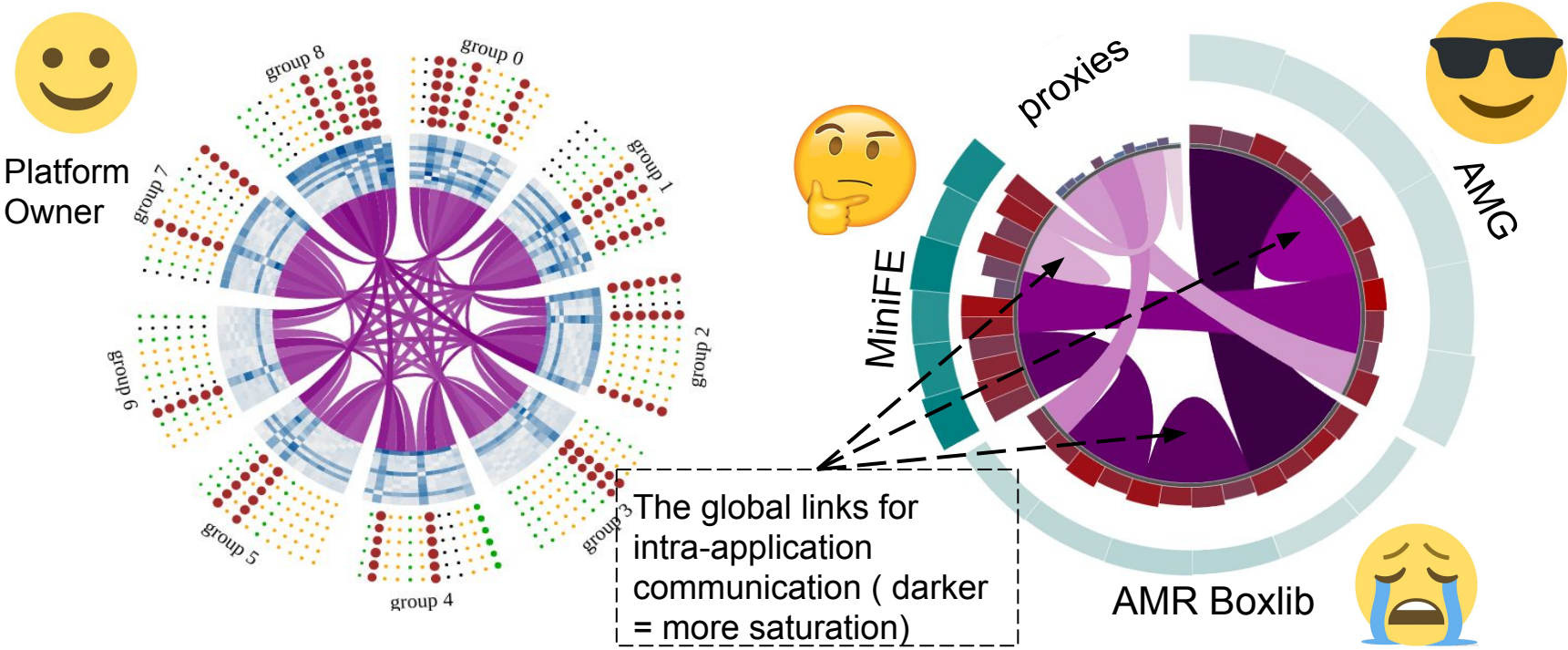
Local link saturation

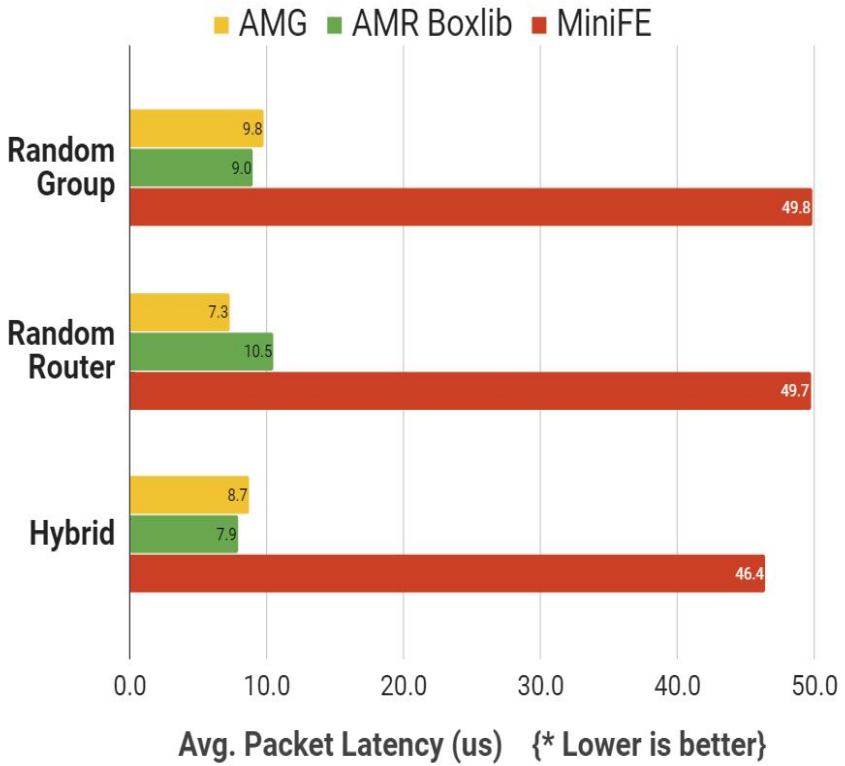
0 1300us

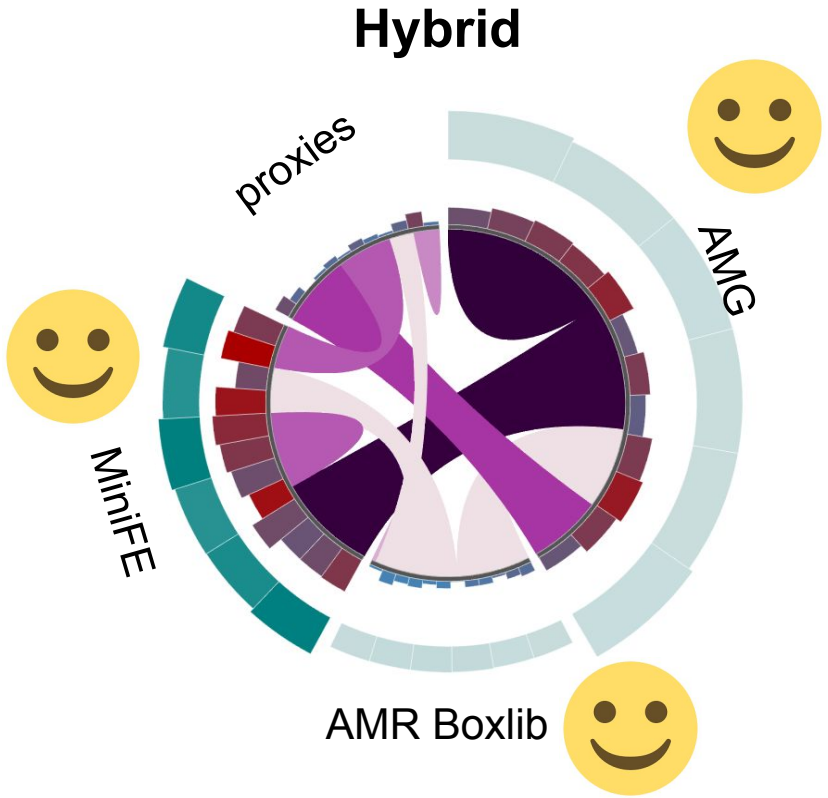
Avg. packet latency

0 54us

Random Router Placement




AMG
AMR Boxlib
MiniFE
Random Group
9.8
9.0
49.8
Random Router
7.3
10.5
49.7
Hybrid
8.7
7.9
46.4
0.0
10.0
20.0
30.0
40.0
50.0
Avg. Packet Latency (us) {* Lower is better}


Hybrid
proxies
AMG
MiniFE
AMR Boxlib


Global link saturation
0
420us


Local link saturation
0
1300us


Avg. packet latency
0
54us

Summary and Future Work

Summary

- Hierarchical aggregation and visualization for HPC networks
- Declarative language for specifying useful visualizations

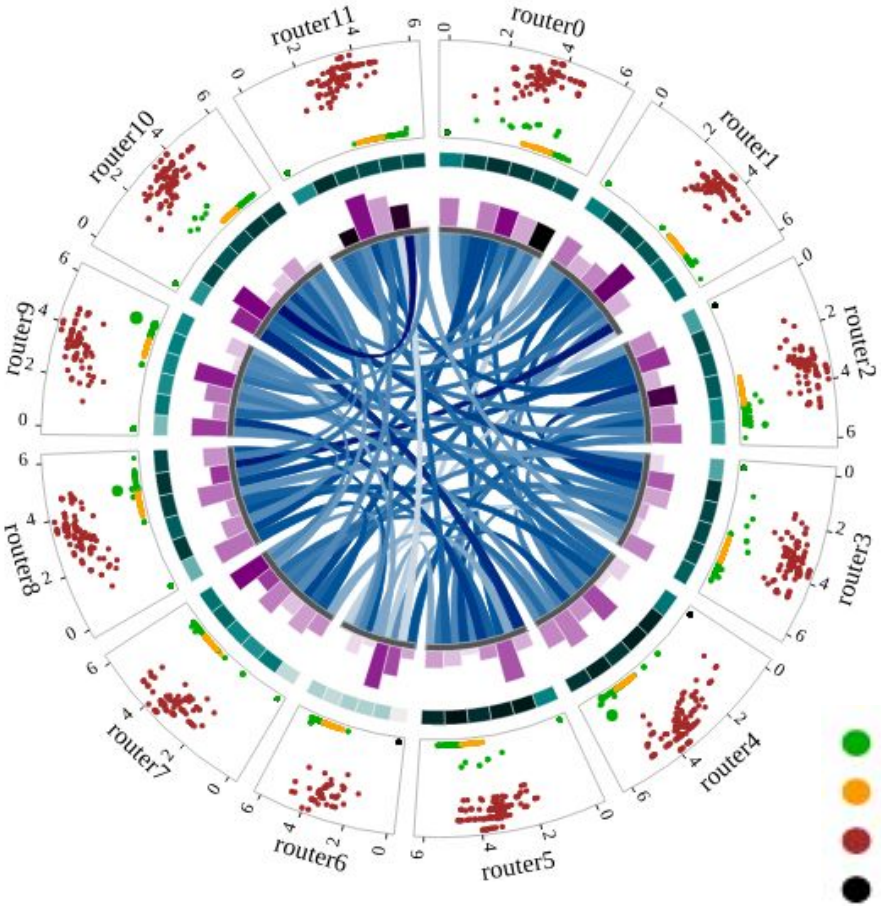
Future Work

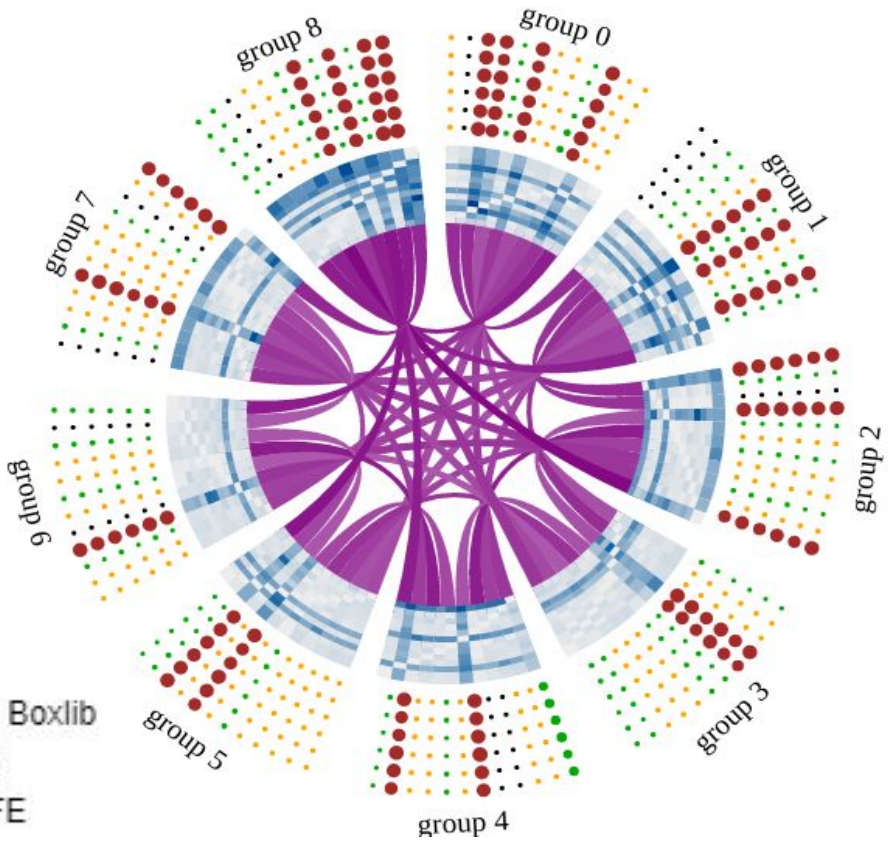
- Other network topologies (e.g., Fat Tree, SlimFly)
- ROSS data for simulator performance analysis
- Cross-domain analysis of model-level data and simulation performance

Thank You!

ACKNOWLEDGMENT

This research has been sponsored in part by the U.S. Department of Energy through grants DE-SC0007443, DESC0012610, and DE-SC0014917. Argonne National Laboratory's work was supported by the U.S. Department of Energy, Office of Science under contract DE-AC02-06CH11357.

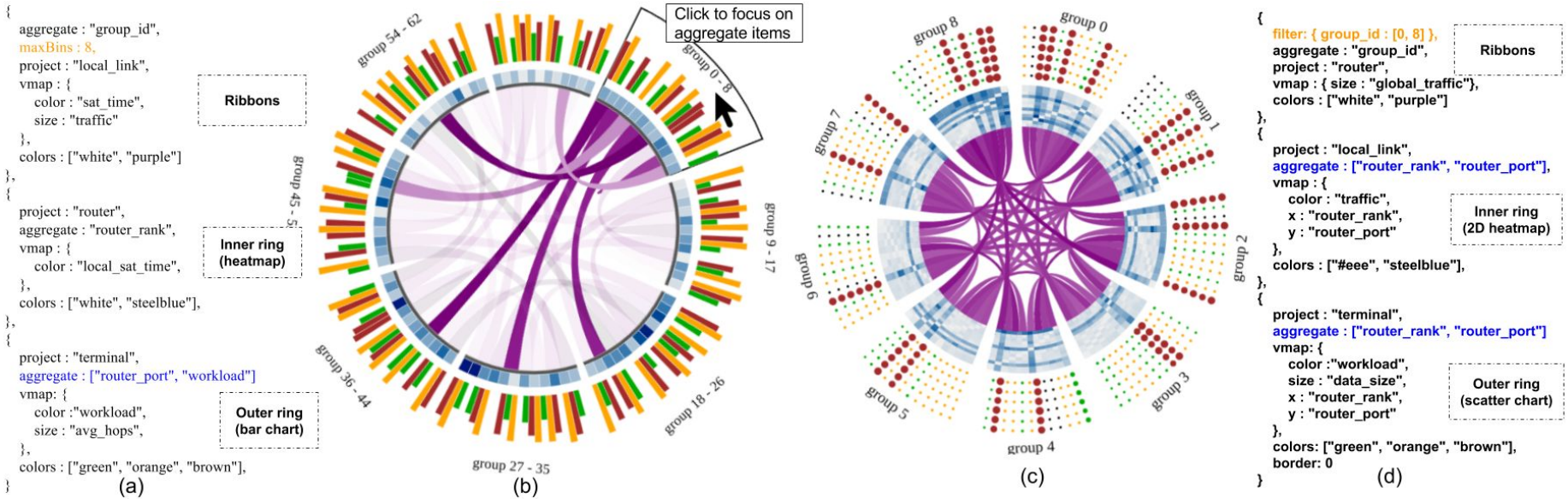




Intra-group Communication

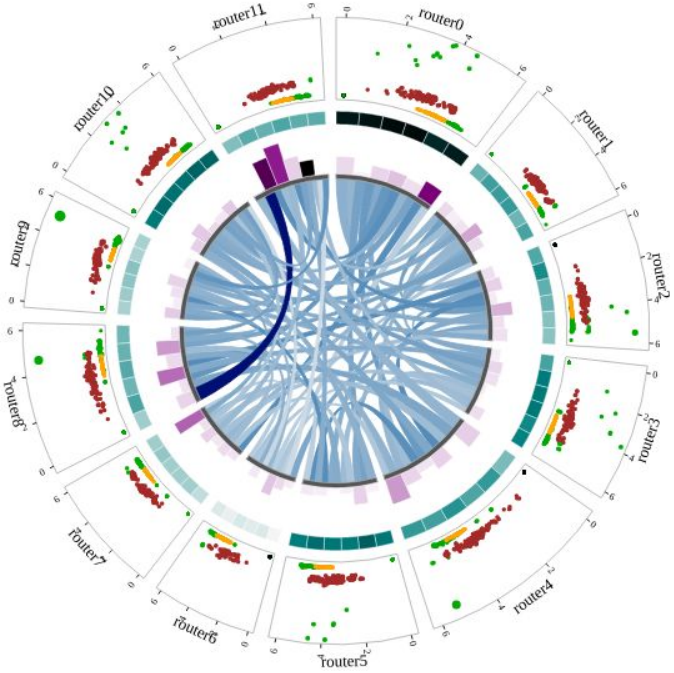
Inter-group Communication

Interactive Projection View



Data Aggregation and Visual Encoding

```
{
    aggregate:  “router_rank”,
    project: “local_link”,
    vmap: {
        color: “sat_time”,
        size: “traffic”
    },
    colors: [“white”, “blue”]
    ….
}
```



Analyzing Time Series Data