ABHINAV BHATELE :: TODD GAMBLIN :: STEVEN H. LANGER :: PEER-TIMO BREMER :: MARTIN SCHULZ

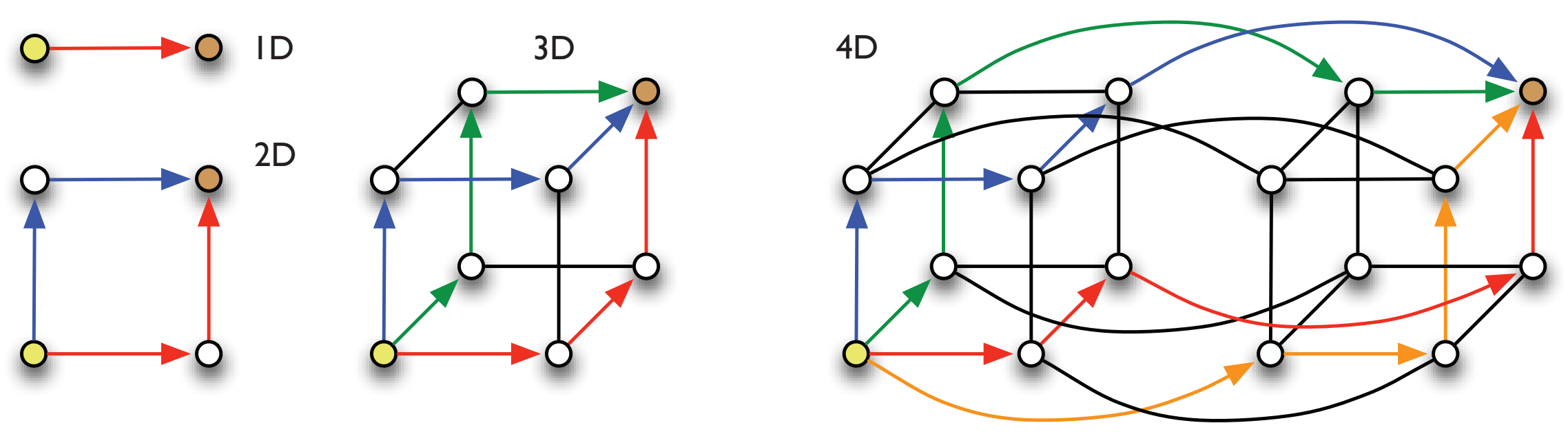
Mapping collectives over sub-communicators on torus networks

Center for Applied
Scientific Computing,
Lawrence Livermore
National Laboratory

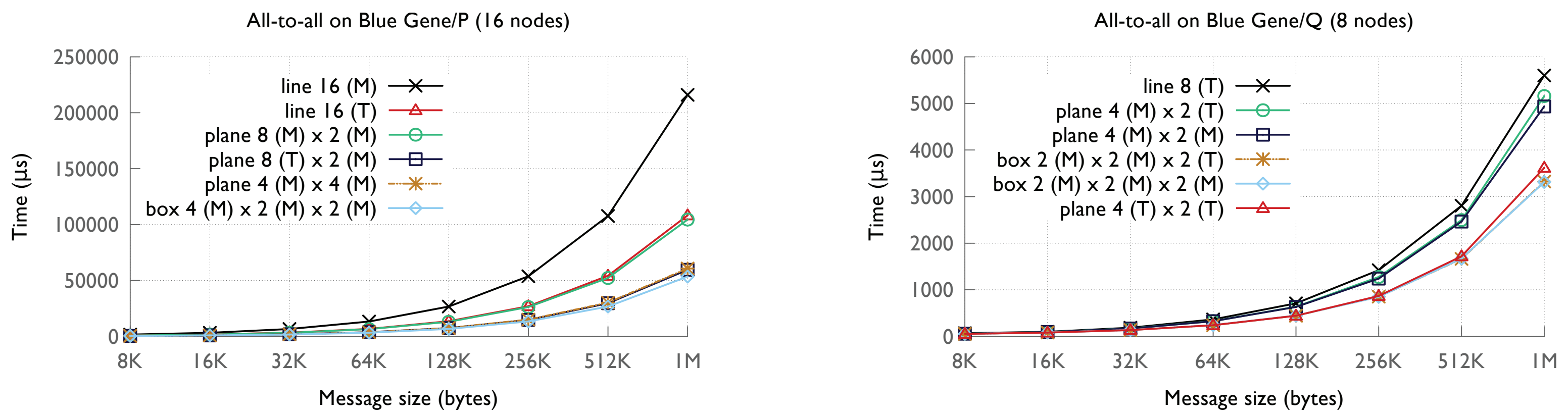
scalability.llnl.gov

The placement of tasks in a parallel application on specific nodes of a supercomputer can significantly impact performance. Traditionally, this task mapping has focused on reducing the distance between communicating processes on the physical network. This minimizes the number of hops that point-to-point messages travel and thus reduces link sharing between messages and contention. However, for applications that use collectives over sub-communicators, this strategy may not be optimal. Many collectives can benefit from an increase in bandwidth even at the cost of an increase in hop count, especially when sending large messages. For example, placing communicating processes in a cube configuration rather than a plane or a line within the network topology increases the number of possible paths messages might take. This increases the available bandwidth which can lead to significant performance gains. We have developed a tool, Rubik, that provides a simple and intuitive API to create a wide variety of mappings for structured communication patterns. Using Rubik, we demonstrate the use of bandwidth maximization techniques for improving the performance of one of the NIF codes, pF3D, which uses collectives over sub-communicators.

Bandwidth optimizations



The location of communicating pairs on an n-dimensional mesh topology determines the number of paths available for message routing. Placing communicating pairs at the diagonally opposite corners of a hypercube provides the highest available bandwidth.

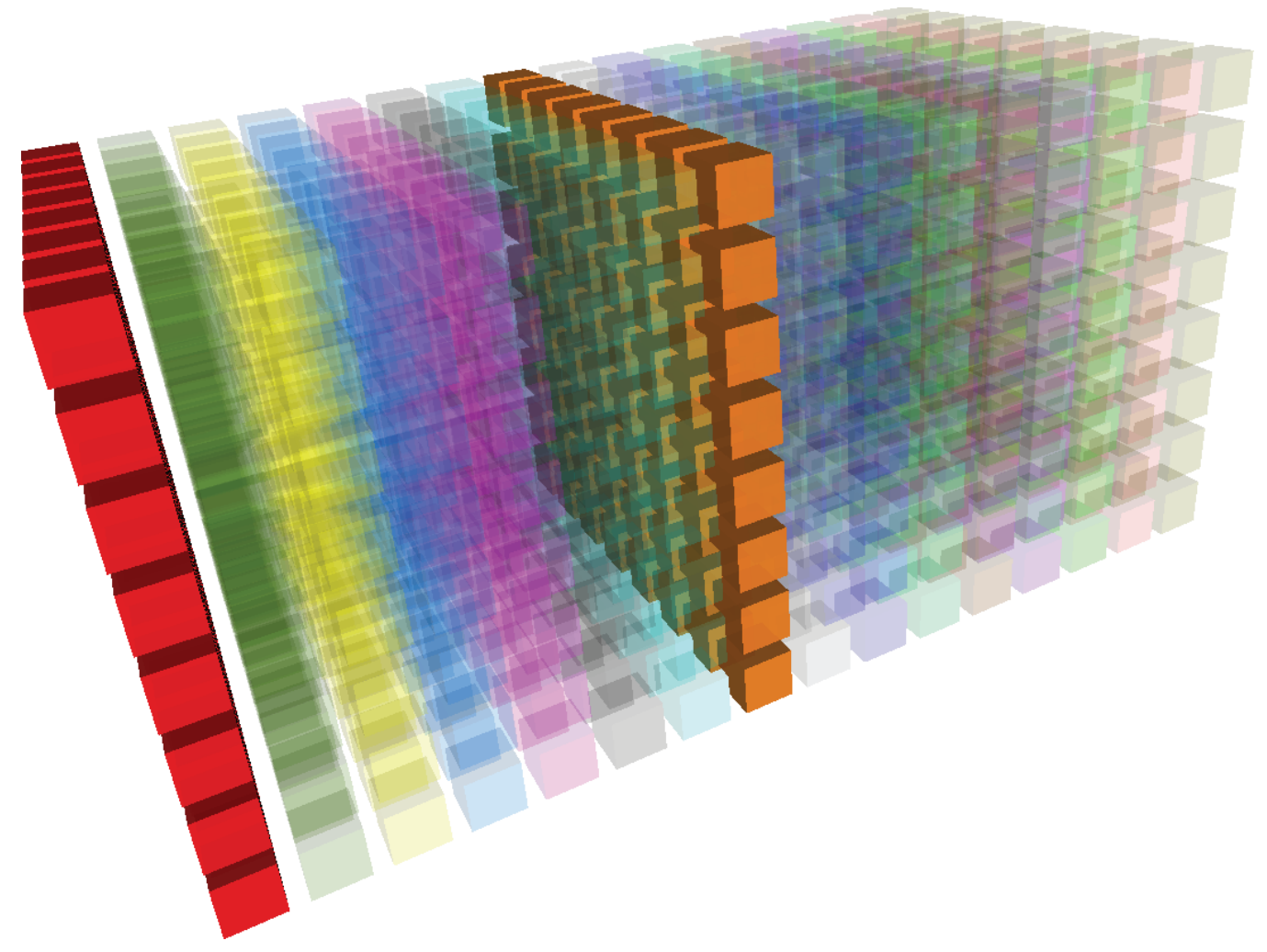


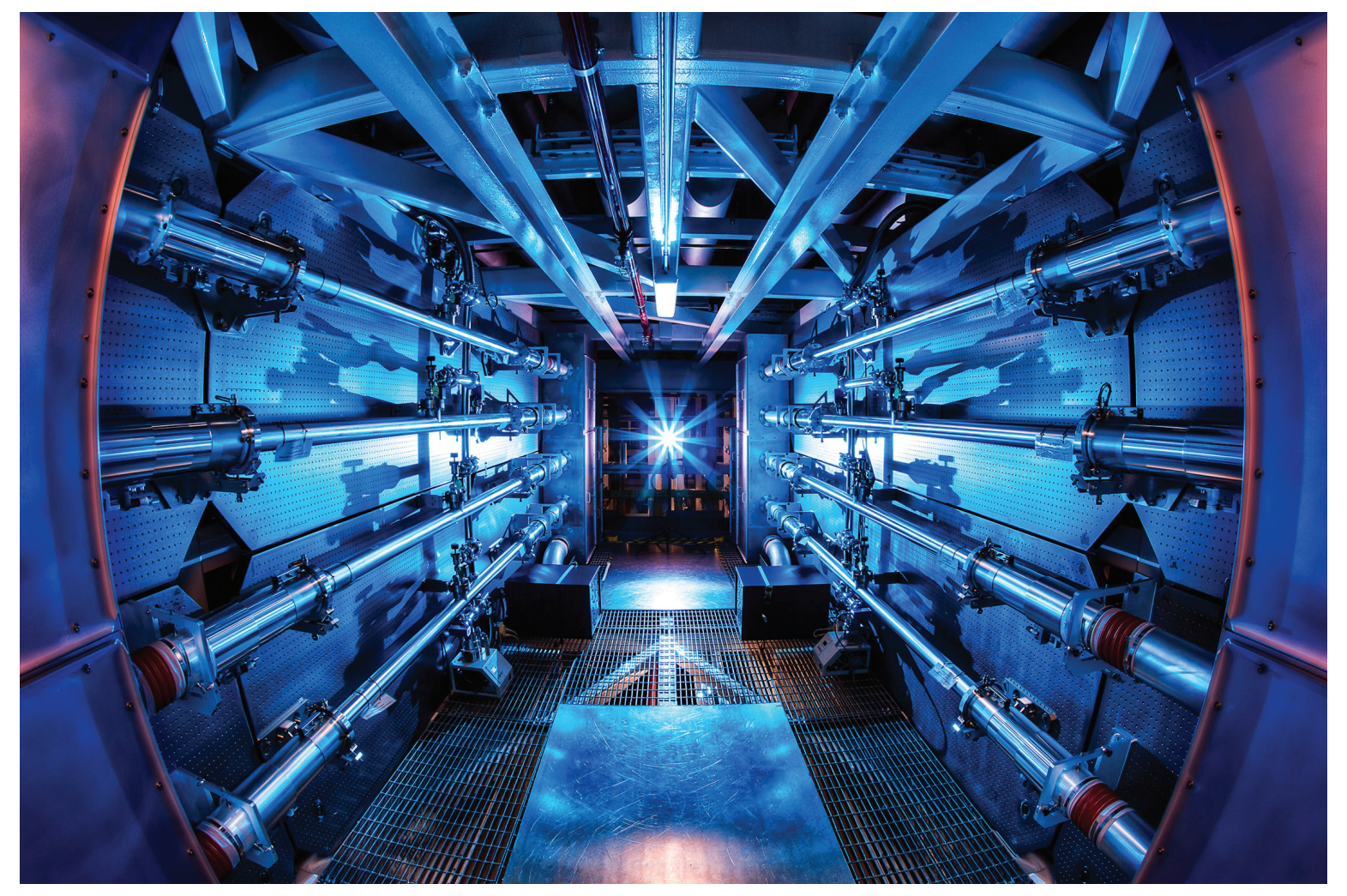
Plots showing the change in performance of an MPI all-to-all depending on the shape of the communicator (Blue Gene/P, left and Blue Gene/Q, right).

pF3D: laser-plasma interaction code

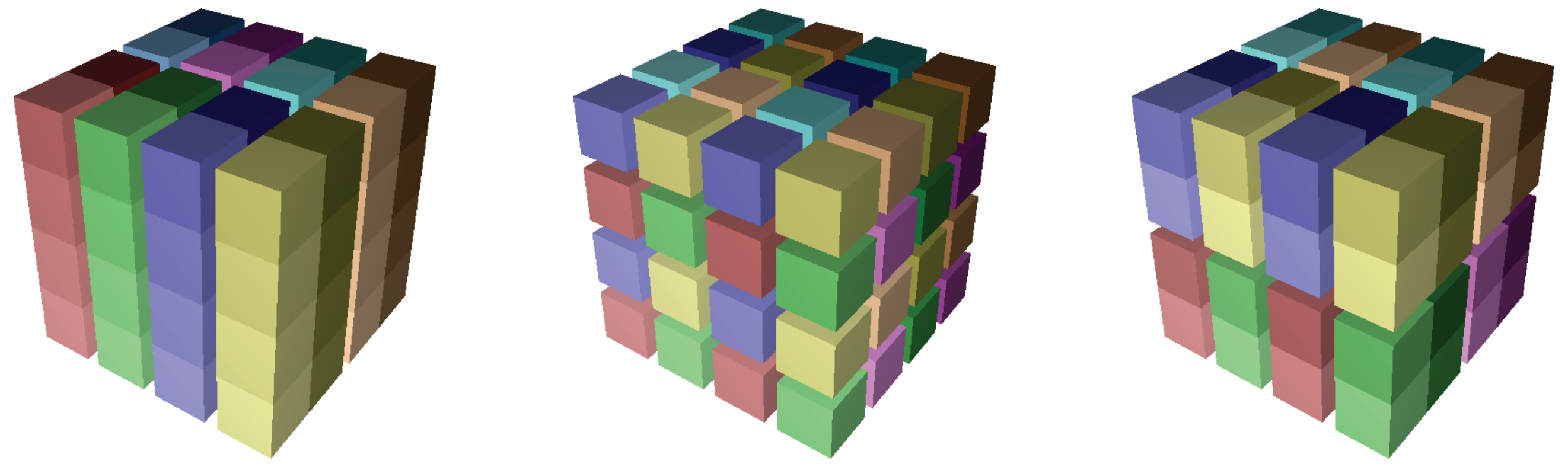
pF3D is a multi-physics code used to study laser-plasma interactions in experiments conducted at NIF. One-dimensional all-to-alls are performed over X and Y communicators within each plane in a three-dimensional virtual process topology.

The preamplifiers of NIF are the first step in increasing the energy of laser beams as they make their way toward the target chamber. NIF recently achieved a 500 terawatt shot - 1,000 times more power than the United States uses at any instant in time.

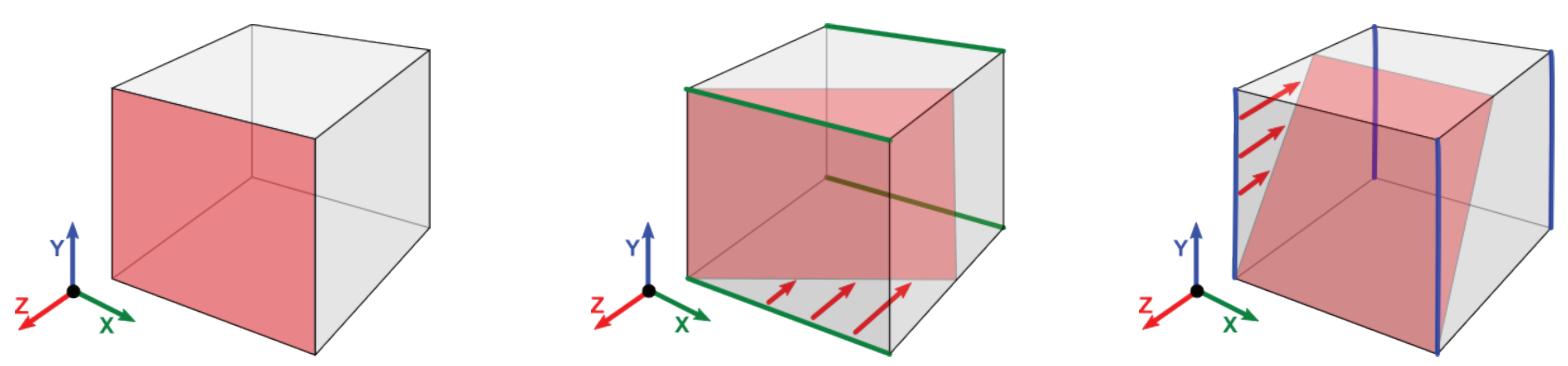




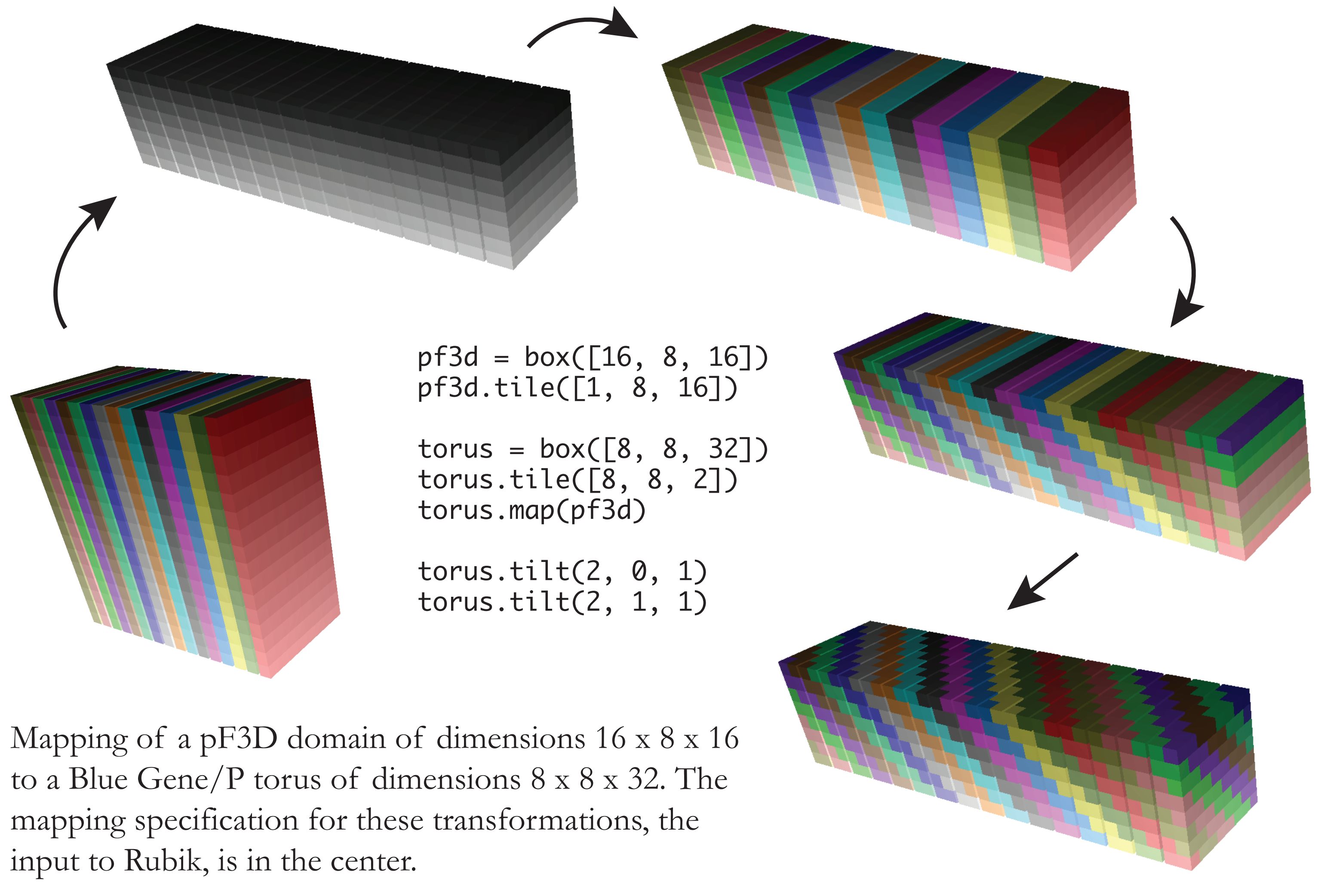
Structured mapping using Rubik



Rubik is a tool for generating mappings of Cartesian virtual topologies on n-dimensional mesh networks. The figures above show the partitioning operations supported by Rubik to divide the application tasks or processors into groups: div or tile (left), mod (center), and cut (right).



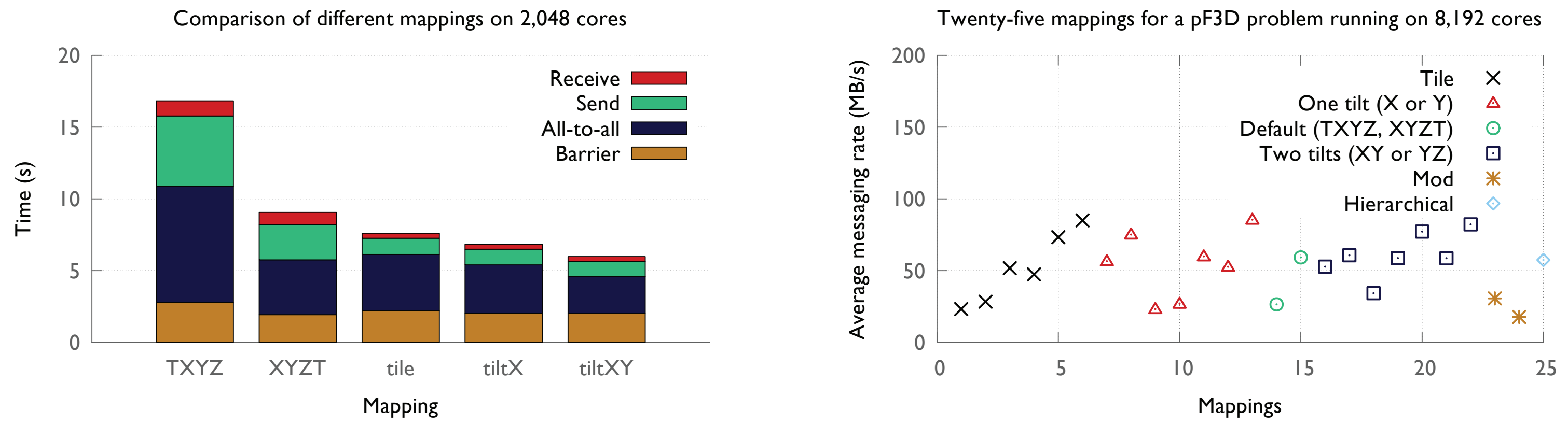
Tilting can be used to increase the bandwidth utilization by moving communicating tasks to diagonally opposite corners. The figures above show an untilted XY plane (left), an XY plane tilted along X (center), and an XY plane tilted along Y (right).



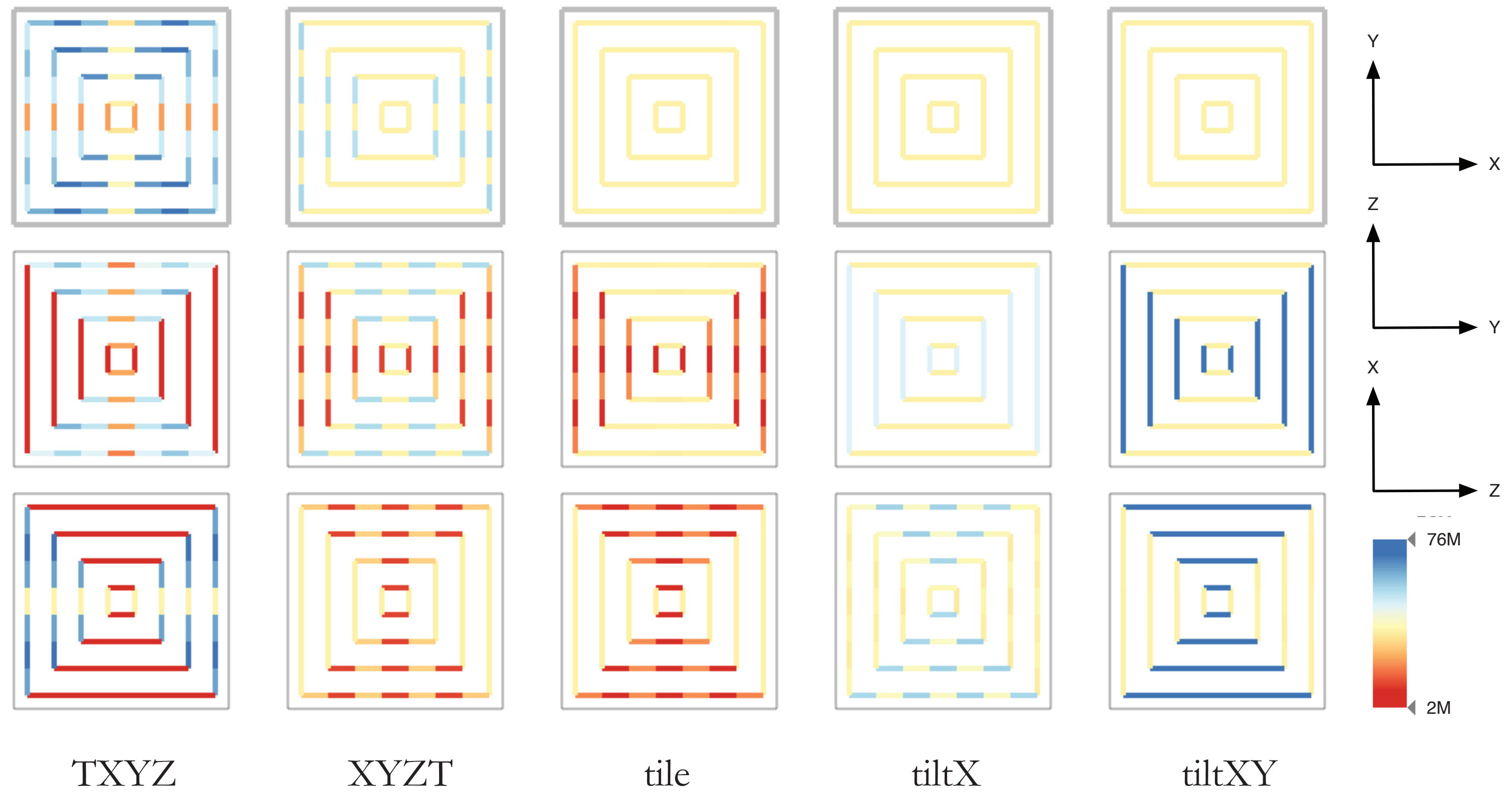
```
pf3d = box([16, 8, 16])
pf3d.tile([1, 8, 16])

torus = box([8, 8, 32])
torus.tile([8, 8, 2])
torus.map(pf3d)

torus.tilt(2, 0, 1)
torus.tilt(2, 1, 1)
```

Mapping of a pF3D domain of dimensions 16 x 8 x 16 to a Blue Gene/P torus of dimensions 8 x 8 x 32. The mapping specification for these transformations, the input to Rubik, is in the center.

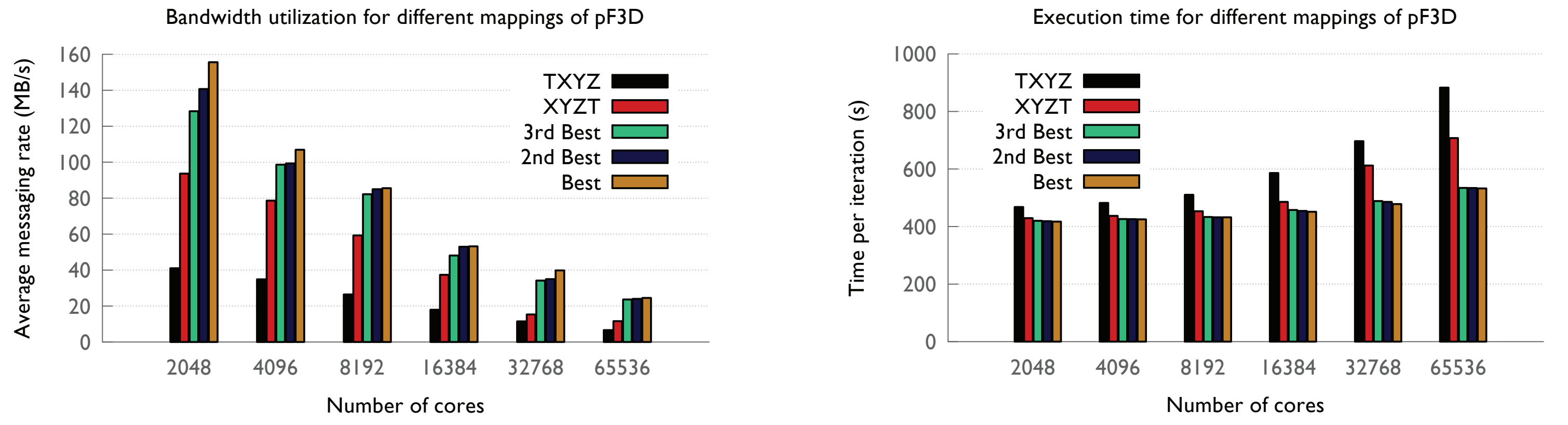
Performance improvements



Plots showing performance improvements for pF3D on 2,048 and 8,192 cores of Blue Gene/P using different mappings generated by Rubik.



Visualizations generated using Boxfish showing the network traffic for different mappings on 2,048 cores of Blue Gene/P. These minimap views show 2D projections of the 8 x 8 x 8 3D torus. Each column shows the X, Y and Z direction traffic for a different mapping.



Plots showing the improvement in messaging rates (left) and time per iteration (right) for pF3D for the three best mappings compared to the default mapping.

Photo credit for the NIF image: Damien Jemison, Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-POST-563791).