Global Arrays Roadmap & Future Developments

Presented by: Daniel Chavarría

GA Team:

Jeff Daily

Sriram Krishnamoorthy

Manoj Krishnan

Bruce Palmer

Abhinav Vishnu

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Global Arrays is used in many scalable applications

- Computational Chemistry
  - NWChem, GAMESS UK, MOLPRO, MOLCASS
- Subsurface Transport Simulation
  - STOMP (recent scaling results on 16k processors)
- Bioinformatics
  - ScalaBLAST
- Computational Fluid Dynamics
  - Tethys

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Global Arrays Successes

- Leading one-sided (put, get) programming model
- Developed since 1992 in conjunction with NWChem
  - Predates MPI (1994)
  - Led by Jarek Nieplocha & Robert Harrison
- Strong research & software contributions in one-sided communication (ARMCI) and distributed array management
  - Targeting large-scale HPC platforms
- Ease of use, productivity and high performance are the key targets for GA
  - GA runtime system enables clean separation of concerns and agility for expansion
- Participation in the broader HPC programming models & communication libraries community
- How are we moving forward?

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

GA is under active development with an aggressive roadmap

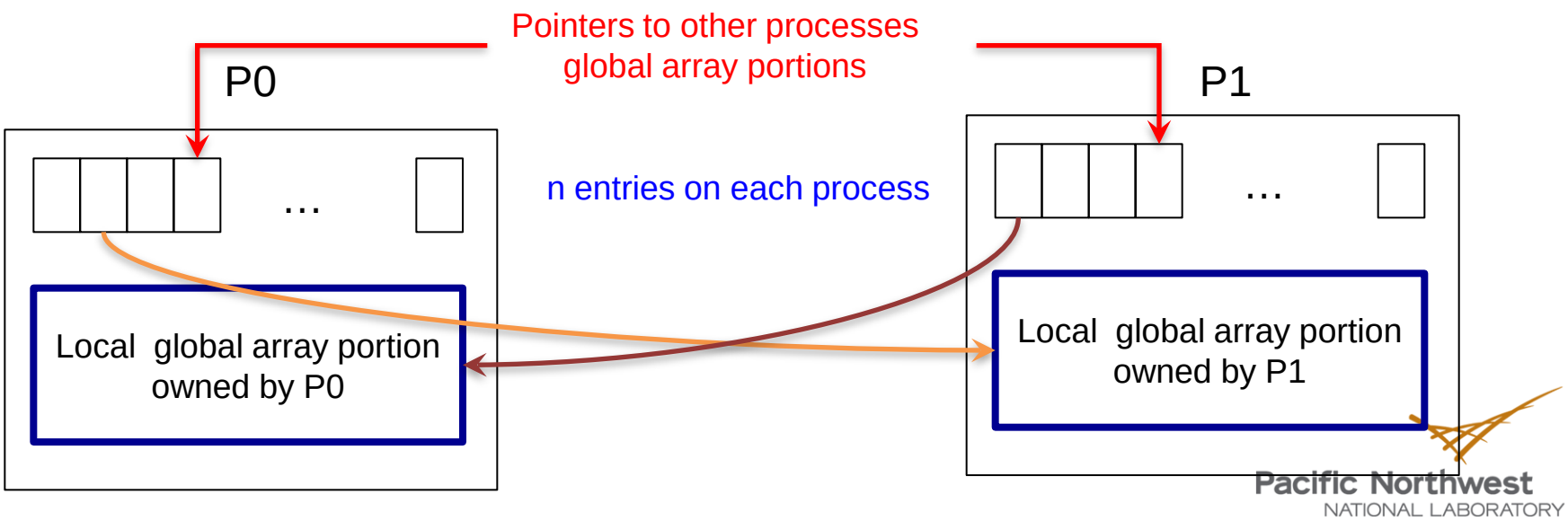
- Scalability to largest available systems (petascale platforms and beyond)
- Portability
- Maintainability and Usability
  - User support
- Fault Tolerance
- Data Decomposition and Load balancing
- Support for Hybrid Platforms
- Performance tools for GA/ARMCI
- Thread safety - support for multithreaded execution
- More GA-based applications

GA has demonstrated scalability on today's largest systems

- Scalability to largest available systems (petascale platforms and beyond)
  - 1.39 PFLOP, 223K process NWChem run using GA on Jaguar was a finalist for the SC'09 Gordon Bell prize (collaboration with ORNL)
- Inter-node and intra-node communication
  - Scaling on manycore, multicore architectures
  - Optimized remote memory access (RMA): put, get, …
  - Prepare for major node architectural shifts
- GA metadata management
  - Reduce memory overhead in GA

Scalability – GA Metadata is a key component

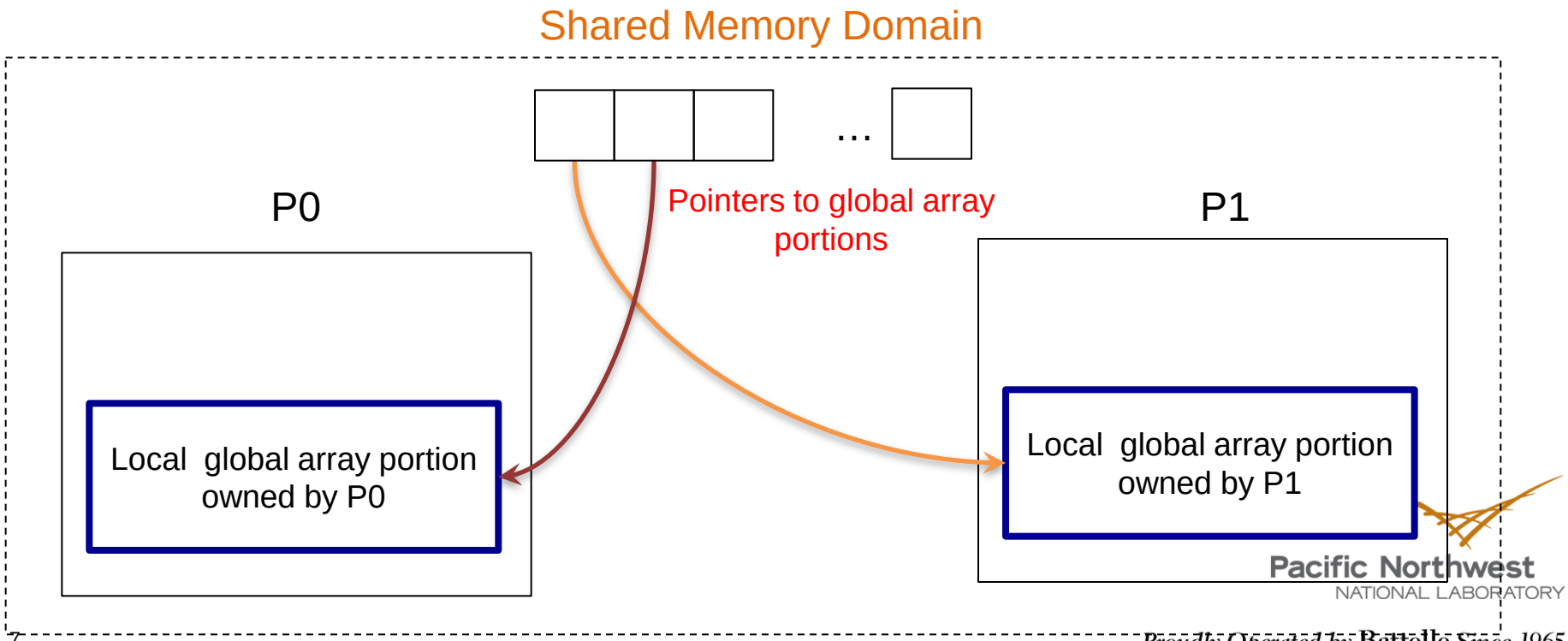
- GA currently allocates metadata for each global array in a replicated manner on each process
- OK for now on petascale systems with $O(10^5)$ processes
  - 200,000 × 8 bytes = 1.5 MB per global array instance
  - Not that many global arrays in a typical application



Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Scalability – Proposed Metadata Overhead Reduction

- Share metadata between processes on the same shared memory domain (today's "node")
- Reduce metadata storage by the number of processes per shared memory domain



Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Portability – GA is available on all large system (current & upcoming)

- Leadership class petaflop machines
  - Cray XTs
  - IBM Bluegene/P
  - (Upcoming) IBM Blue Waters
- Multicore Linux Clusters
  - e.g. Chinook – Infiniband Linux cluster
- Upcoming/future architectures/networks
  - Cray Gemini & Aries, IBM PERCS
  - Hybrid systems
- Maintenance and Usability
  - Modularization of ARMCI protocol layer
  - User support

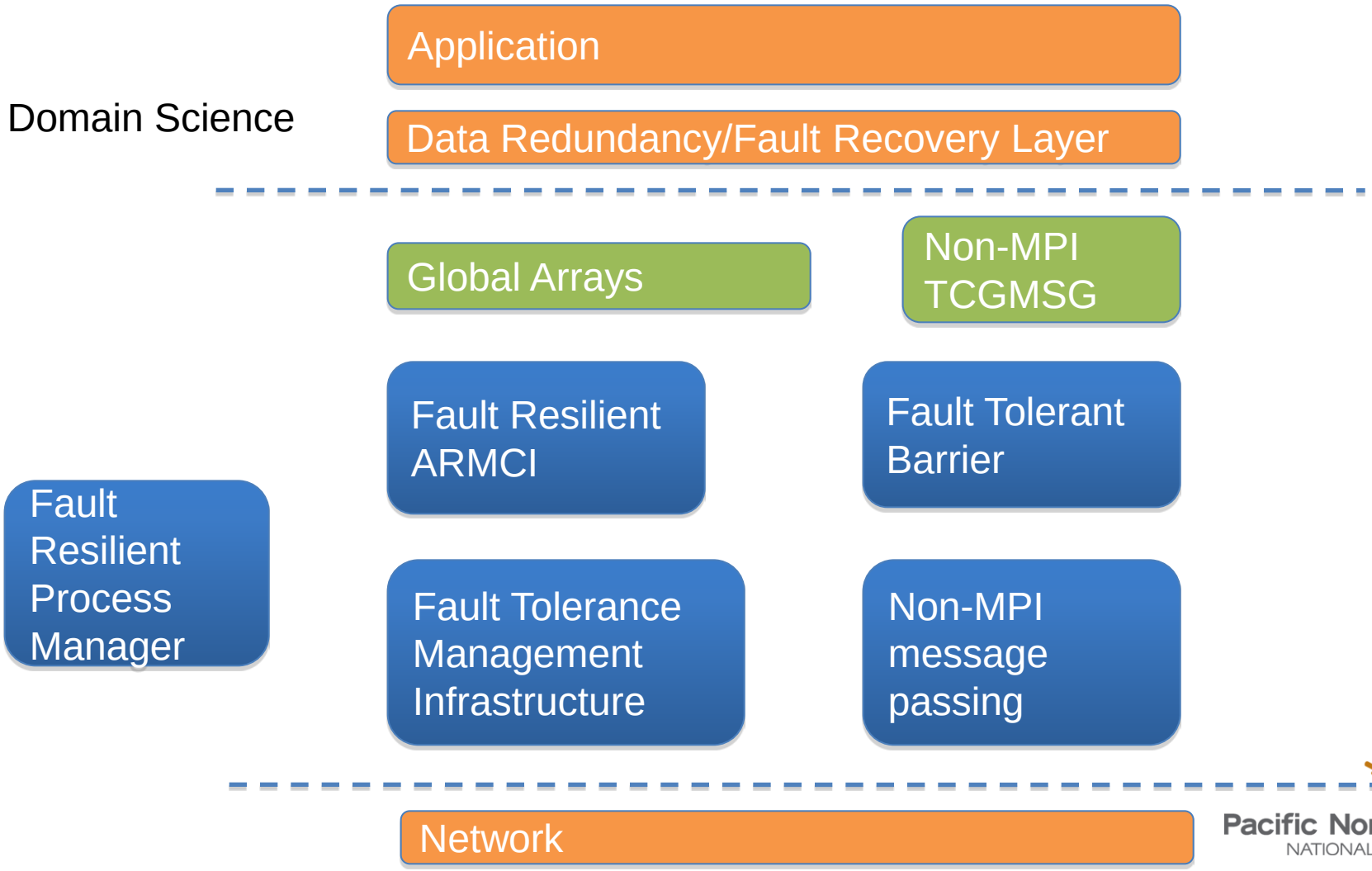
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

GA Extreme Scale Challenges currently being addressed

- Fault Tolerance
  - Fault resilient GA
  - Fault tolerance support to GA applications
- Data Decomposition and Load balancing
  - Flexible and topology-aware data decomposition
  - Dynamic load balancing
    - Task counters, work queues, work stealing, …
  - Communication optimization
- Support for Hybrid Platforms
  - Support GA applications running on GPU-accelerated clusters
- Performance tools for GA/ARMCI

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

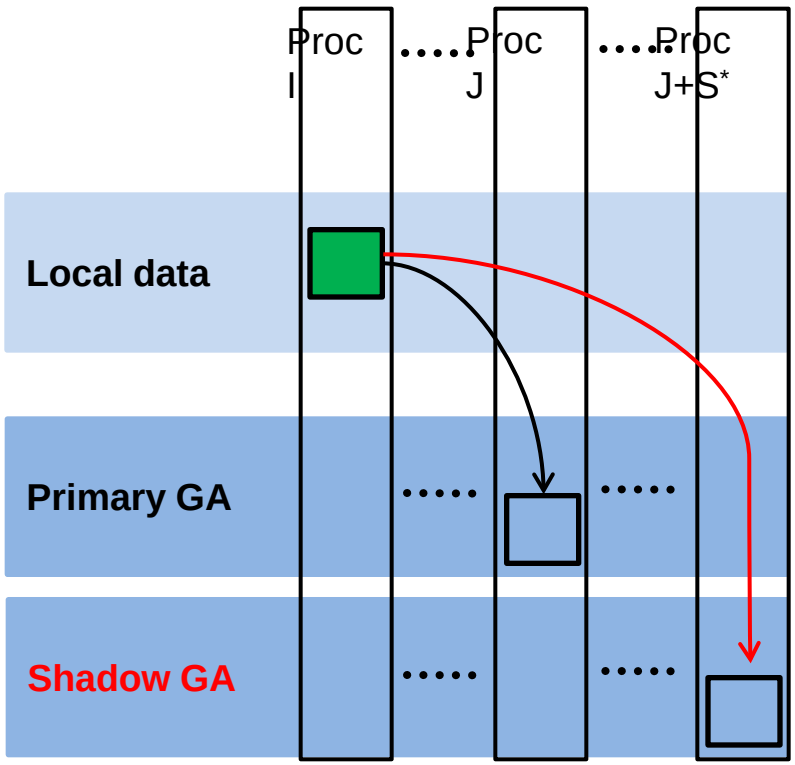
Advanced Fault Tolerance Capabilities



Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Advanced Fault Tolerance Capabilities (cont.)

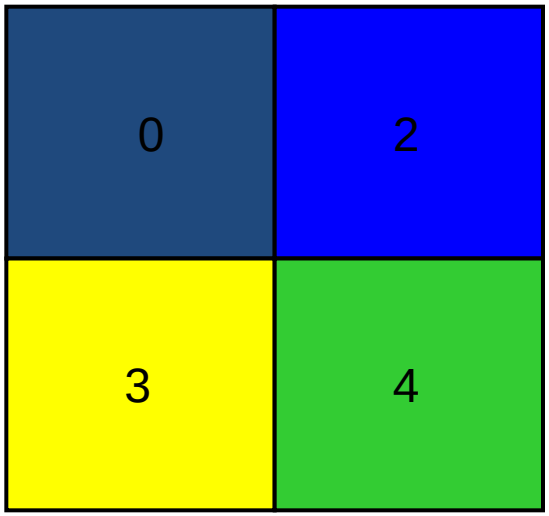
- When process/node fails all data on it becomes inaccessible
- Alternative data source therefore needed
- "S" is the a suitable process shift to ensure safe keeping of data
- Availability of data managed through info from FTMI
- Suitable data access integrated in Get, Put, and Acc operations
- Demonstrated in the context of NWChem modules (see PNNL booth for live demo)

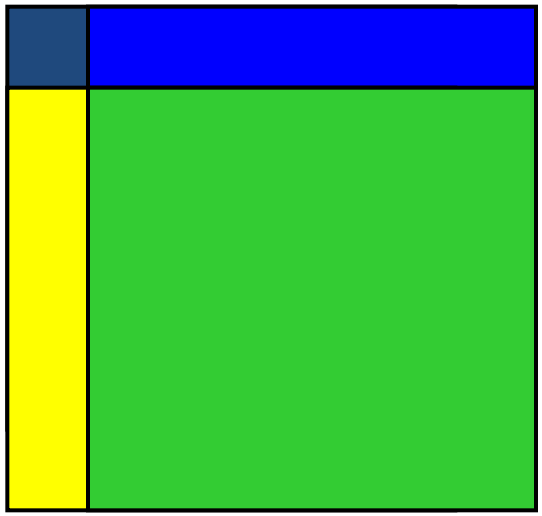


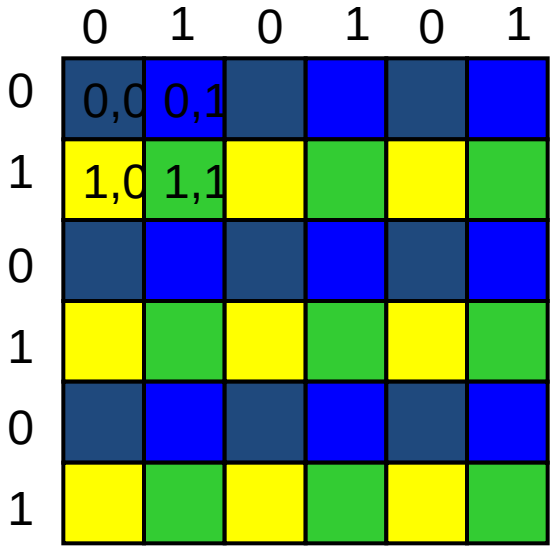
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

GA Supports Common Data Distributions

- Supported distributions
  - Regular, irregular, block cyclic



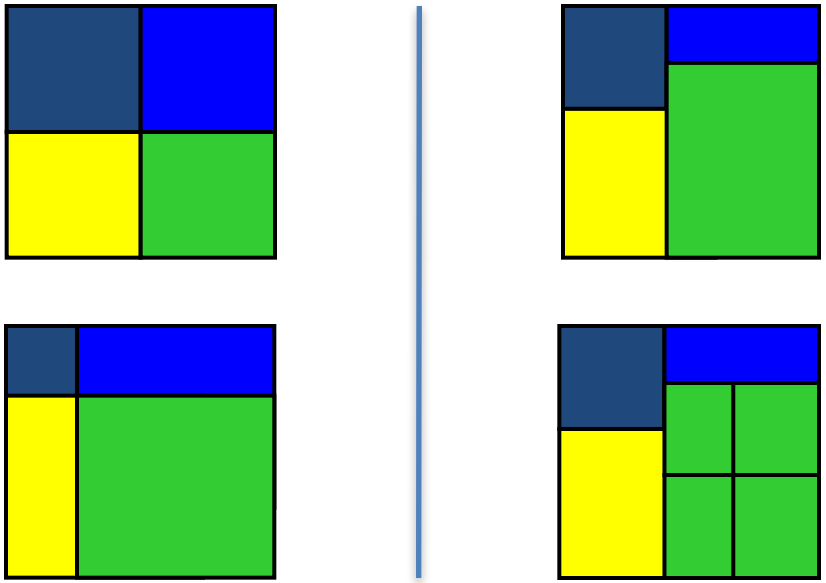


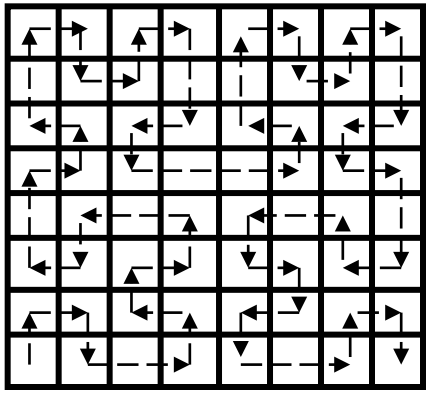


Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Proposed Flexible Data Distributions

- Enable much finer-grained partitioning of data and will provide a basis both for implementing dynamic load balancing schemes and creating layouts that optimize locality and communication
  - Provides complete control for the user to distribute data
  - Supports more than 1 block per process
  - Reduces communication

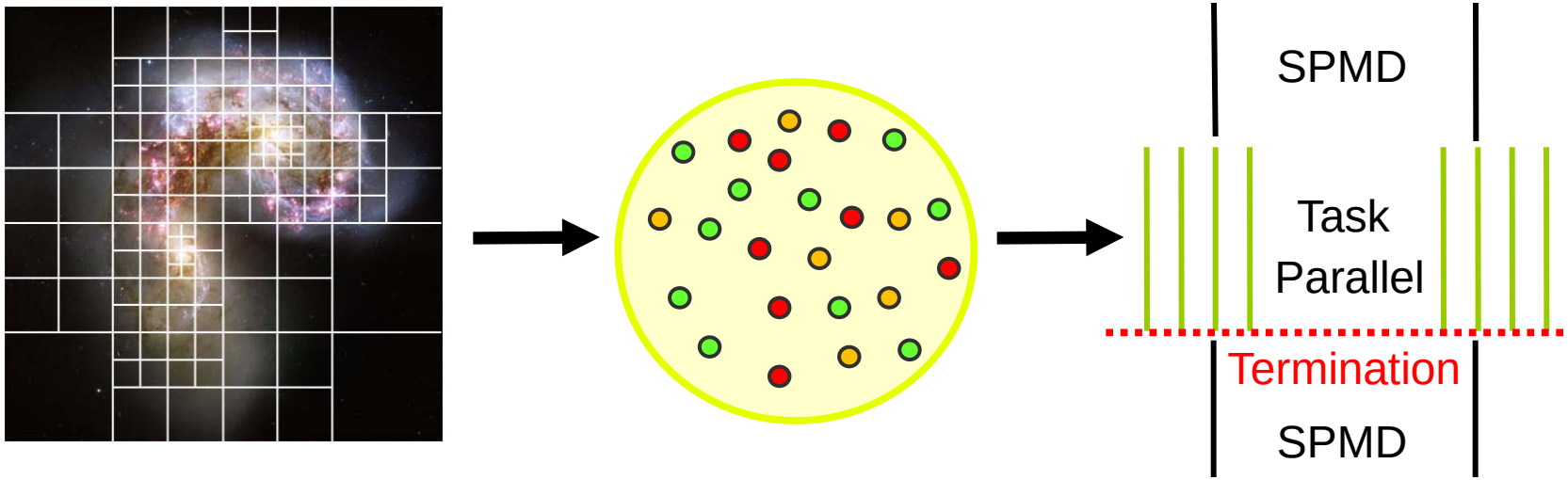




Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

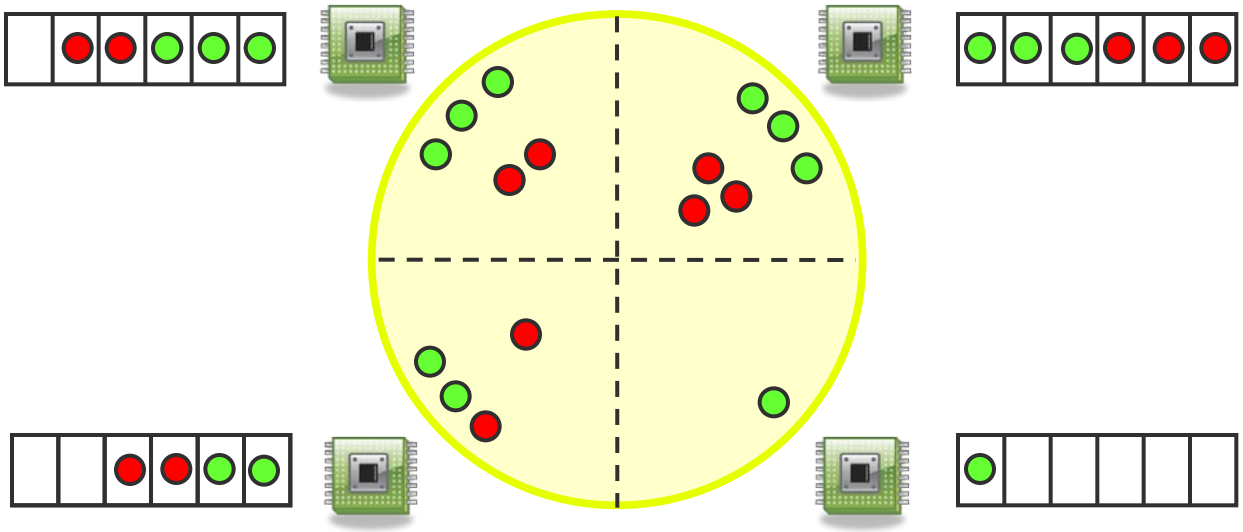
TASCEL –Dynamic Load Balancing



- Express computation as collection of tasks
  - Tasks operate on data stored in PGAS (Global Arrays)
  - Executed in collective task parallel phases
- TASCEL runtime system manages task execution
  - Load balancing, locality optimization, etc.
- Complementary library to Global Arrays

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

TASCEL – Dynamic Load Balancing (cont.)



- *High level interface:* shared global task collection
- *Low level interface*: set of distributed task queues
  - ARMCI task queue on each processor
    - Steals are truly one-sided, don't interrupt remote processor
    - Reactive to load imbalance
    - Distributed scheme – only local information

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Extending Support for thread-based execution models running on multi-core platforms

- Support thread-based execution models running on multi-core platforms
  - OpenMP, `pthreads`, `qthreads`, others
- Allow multiple threads executing on a single node to execute GA function calls to any other node
- Move towards a more multi-core friendly execution model
  - More sharing of memory and other resources compared to pure process-based models

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Working closely with MPI-3 Forum

- The MPI-2 one-sided API – MPI_Get, MPI_Put, etc
  - proposed eleven years ago
  - too restrictive for actual application use
- MPI-3 RMA
  - developing alternative design to the MPI-2 one-sided standard
  - Inspired by ARMCI

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Current release - GA 5.0

- Scalability on leadership class machines
  - Cray XT5, Bluegene/P, large-scale Linux clusters
- Usability
  - GA Install and setup
  - Revamping GA webpage and documentation
- Autoconf/automake build mechanism
- New Python interface

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Summary: GA has a proven track record with an active development path

- The GA paradigm has proven very successful
  - efficient & scalable on a wide range of architectures
    - core operations tuned for high performance
  - library substantially extended but all original (1994) APIs preserved
  - increasing number of application areas
- Supported and portable tool that works in real applications

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Moving Forward

- Very strong support from PNNL through the eXtreme Scale Computing Initiative (XSCI)
  - Core CS focus on Global Arrays
- Continued support & investment from DOE
- New GA capabilities & software releases coming online
- Strong participation in community efforts to support unified runtime systems
  - unistack effort
- Actively seeking further collaborations to guide future capabilities and application support

Source Code and More Information

- Version 5.0 available
- Homepage at http://www.emsl.pnl.gov/docs/global/
- Platforms (32 and 64 bit)
  - IBM BlueGene/L, BlueGene/P, LAPI-based systems
  - Cray XTs
  - Linux clusters with InfiniBand, Ethernet, Myrinet, or Quadrics
  - Other systems (SGI Altix, Fujitsu, NEC, Windows)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965