EPiGRAM: Preparing Message-Passing and PGAS Programming Models for Exascale

Stefano Markidis and Erwin Laure
KTH Royal Institute of Technology, Stockholm, Sweden
Exascale Programming Models

• With the evolution of HPC architecture towards exascale, new approaches for programming these machines need to be found
  – EPiGRAM focuses on exploring programming models for the exascale era.
• Intense discussion on
  – existing programming models can be improved to exascale with an incremental MPI-like approach.
  – Disruptive changes are needed.
Plan A to Exascale

Devise a new programming model
  – Ideally high level to increase productivity
  – Including autotuning and adaptivity
  – Deals efficiently with heterogeneous hardware
    • Combination of compiler/runtime system

These are important research questions one should (and people actually do) work on
  – But will take a long time before usable in real applications. Ready for 2024?
Plan B to Exascale

Work on improving existing, widely used programming systems:

– MPI
– OpenMP
– Recently PGAS has also gained momentum
– Cuda/OpenCL/OpenACC

**New developments** in MPI (3.1) and OpenMP (4) gained strong momentum. Need to maintain it to reach exascale
EPiGRAM Key Players

- **KTH**: management and applications
- **TUW**: exascale MPI
- **FRAUNHOFER**: exascale PGAS
- **CRAY UK**: programming models for diverse memory spaces
- **EPCC**: PGAS-based MPI
- External Contributor: **UNIVERSITY OF ILLINOIS**: exascale MPI
EPiGRAM Focuses on MP and PGAS

– Proven petascale technologies
– MPI still the most widely used
– MPI and GASPI (GPI) programming systems
– New disruptive programming approaches use MPI and PGAS-like transport layers at lower levels.
EPiGRAM AT A GLANCE
EXASCALE PARALLELISM

MEMORY CONSUMPTION
COLLECTIVE OPERATIONS
EMERGING COMPUTING MODELS

INTEROPERABILITY

MPI ZERO-COPY COLL.
GPI DYNAMIC CONN.
MPI ISOMORPHIC COLL.
GPI NON-BLOCK. COLL.
MPI STREAMS

PGAS-BASED MPI
GPI INTEROPERABILITY WITH MPI
MPI ENDPOINTS

STANDARDIZATION: GASPI FORUM AND MPI FORUM

APPLICATIONS
iPIC3D, Nek5000 and IFS
Memory Consumption

1. MPI Dealing with limited and slower memory:
   - in-depth analysis of MPI derived datatype mechanism for saving copy-operations;
   - Space efficient representation of derived datatypes.
   - Advanced datatype library available at http://www.epigram-project.eu/advanced-datatype-library/
   - analysis of MPI collective interface specification with suggestions for improvement at http://www.epigram-project.eu/zero-copy-all-to-all-communication-with-mpi-datatypes/

2. GPI Dynamics connections
   - GPI connections are dynamically established as the first communication request between two nodes is performed. Memory consumption is decreased by a factor of hundreds.
Collective Operations

1. MPI Collective communication at scale:
   – proposal for specification of homogeneous stencils, towards improved (homogeneous, regular) sparse collectives

2. Non-blocking GPI collectives
   – Increased scalability of collective operations and synchronization in GASPI/GPI
EXASCALE PARALLELISM

EMERGING COMPUTING MODELS

MPI STREAMS
MPI Streams

- Designed and implemented as an extension to MPI to support streaming computing models
- Showed that can support high-processing rates as required by SKA, LSST, LHC ATLAS.
- Deployed in HPC applications, such as iPIC3D and CG solver.
PGAS-BASED MPI: EMPI4Re

• Combining MP and PGAS in one MPI implementation, called EMPI4Re
  – MPI 1.0 atop:
    • DMAPP (CRAY)
    • GPI (for all systems)
  – Research vehicle for investigating new concepts in MPI, such as MPI endpoints, persistent collectives, …
GPI-MPI Interoperability and MPI Endpoints

• Increased interoperability by isolating GPI: no need for guarding GPI and MPI epochs with barriers.

• First Implementation of MPI Endpoints in EMPI4Re (current proposal in MPI Forum).

Euro-MPI 2013; Courtesy Jim Dinan
EXASCALE PARALLELISM

MEMORY CONSUMPTION

COLLECTIVE OPERATIONS

EMERGING COMPUTING MODELS

MPI ZERO-COPY COLL.

GPI DYNAMIC CONN.

MPI ISOMORPHIC COLL.

GPI NON-BLOCK. COLL.

MPI STREAMS

PGAS-BASED MPI

GPI INTEROPERABILITY WITH MPI

MPI ENDPOINTS

APPLICATIONS

iPIC3D, Nek5000 and IFS

INTEROPERABILITY

EPiGRAM
EXASCALE PARALLELISM

MEMORY CONSUMPTION

COLLECTIVE OPERATIONS

EMERGING COMPUTING MODELS

INTEROPERABILITY

MPI ZERO-COPY COLL.

GPI DYNAMIC CONN.

MPI ISOMORPHIC COLL.

GPI NON-BLOCK. COLL.

MPI STREAMS

PGAS-BASED MPI

GPI INTEROPERABILITY WITH MPI

MPI ENDPOINTS

STANDARDIZATION: GASPI FORUM AND MPI FORUM

APPLICATIONS

iPIC3D, Nek5000 and IFS
Standardization

• One of the drivers to establish GASPI-Forum. New proposals tested in EPIGRAM:
  – Application-provided memory segment
  – …

• Present in all MPI-Forum meetings
  – MPI endpoints proposal
  – Persistent collective proposal
  – …
Summary

• Exascale technologies pose new challenges on efficient programming
  – Hybrid forms of parallelism, deep and limited memory hierarchies, dynamic behavior, etc.
• Evolutionary path that builds on existing, widely used components is likely more promising than designing a new programming model
• MPI and PGAS are good starting points for such an evolutionary path
  – Interoperability is a key issue!
  – Higher abstractions potentially disruptive can (should) be built on

The EPiGRAM Project: Preparing Parallel Programming Models for Exascale

Stefano Markidis, Ivy Bo Peng, Jesper Larsson Träff, Antoine Rougier, Valeria Bartsch, Rui Machado, Mirko Rahn, Alistair Hart, Daniel Holmes and 2 more

http://link.springer.com/chapter/10.1007/978-3-319-46079-6_5
Thank you!

http://www.epigram-project.eu/