IHK/McKernel: A Lightweight Multi-kernel Operating System for Extreme-Scale Supercomputing

Balazs Gerofi
Exascale System Software Team,
RIKEN Center for Computational Science

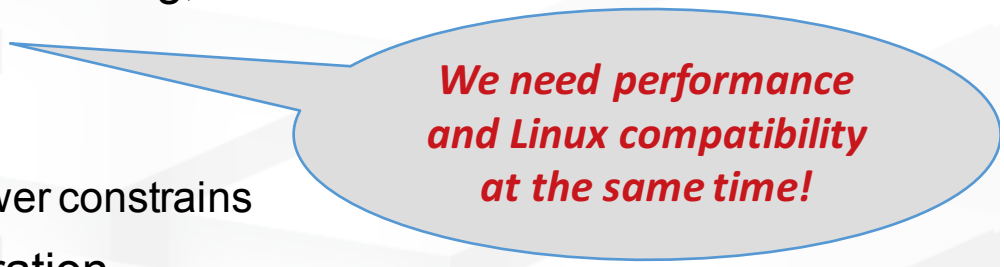
2018/Nov/15 – SC'18 Intel Extreme Computing Users Group (IXPUG) BoF

RIKEN

Motivation – System software/OS challenges for high-end HPC

- **Node architecture: increasing complexity**
  - Large number of (possibly heterogeneous) processing cores, deep memory hierarchy, complex cache/NUMA topology
- **Applications: increasing diversity**
  - Traditional/regular HPC + in-situ data analytics + Big Data processing + AI / Machine Learning + Workflows, etc.

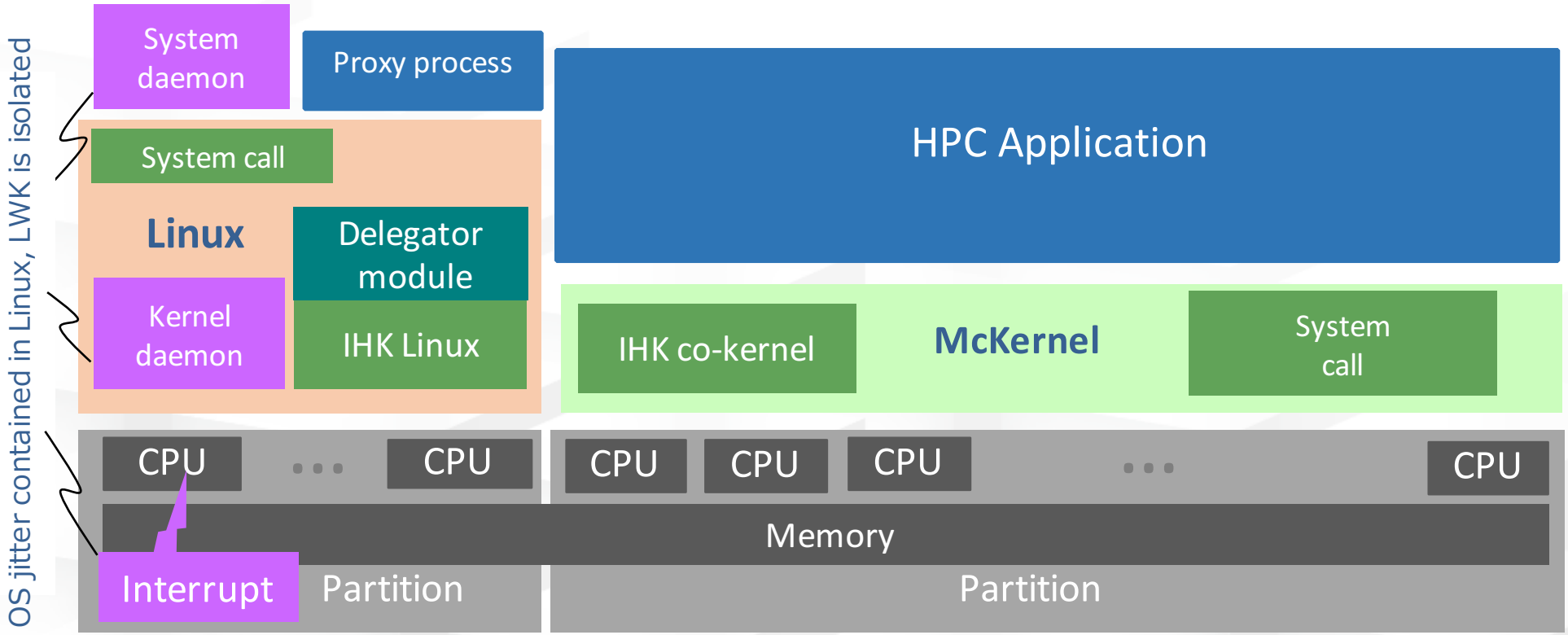
- **What do we need from the system software/OS?**
  - Performance and scalability for large scale parallel apps
  - Support for Linux APIs – tools, productivity, monitoring, etc.
  - Full control over HW resources
  - Ability to adapt to HW changes
    - Emerging memory technologies, parallelism, power constrains
  - Performance isolation and dynamic reconfiguration
    - According to workload characteristics, support for co-location



RIKEN

IHK/McKernel: Lightweight Multi-kernel Architecture

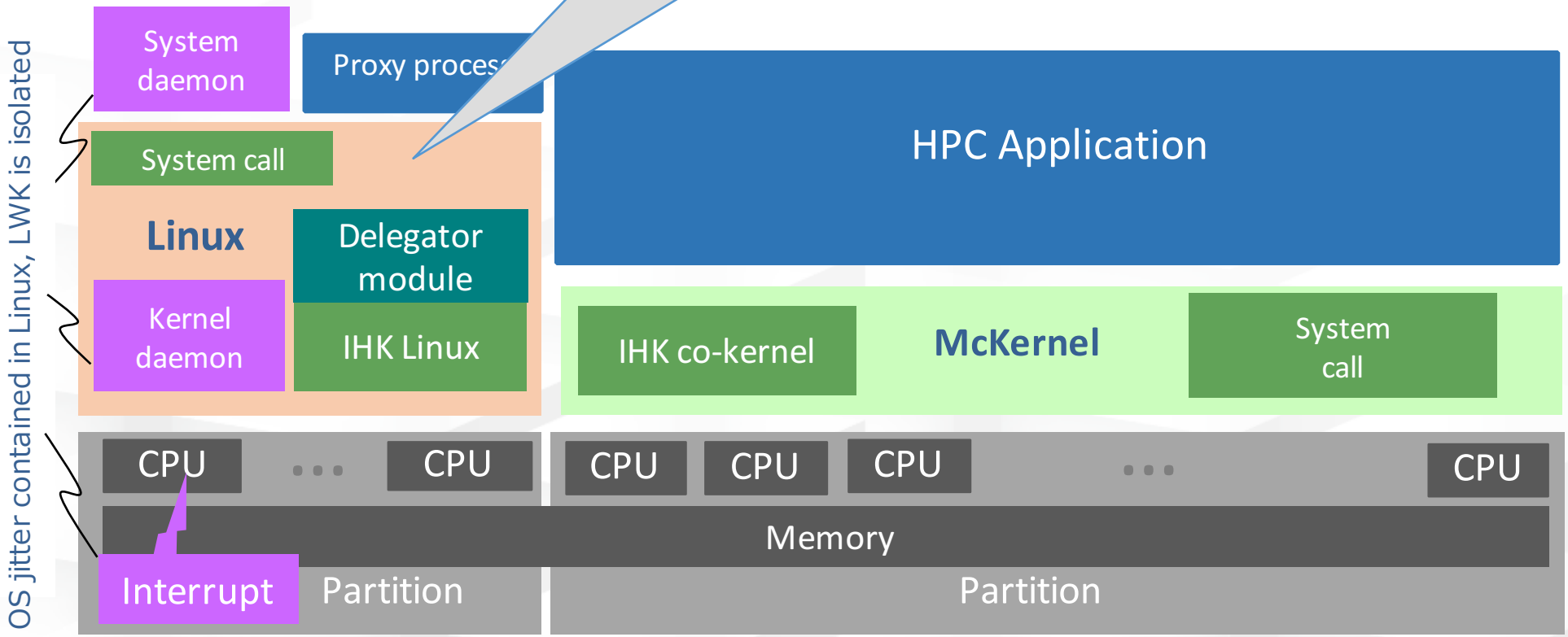
- **Interface for Heterogeneous Kernels (IHK):**
  - Allows *dynamic partitioning* of node resources (i.e., CPU cores, physical memory, etc.)
  - Enables management of multi-kernels (assign resources, load, boot, destroy, etc..)
  - Provides inter-kernel communication (IKC), messaging and notification
- **McKernel:**
  - A lightweight kernel developed from scratch, boots from IHK
  - Designed for HPC, noiseless, simple, implements only performance sensitive system calls (roughly process and memory management) and the rest are offloaded to Linux



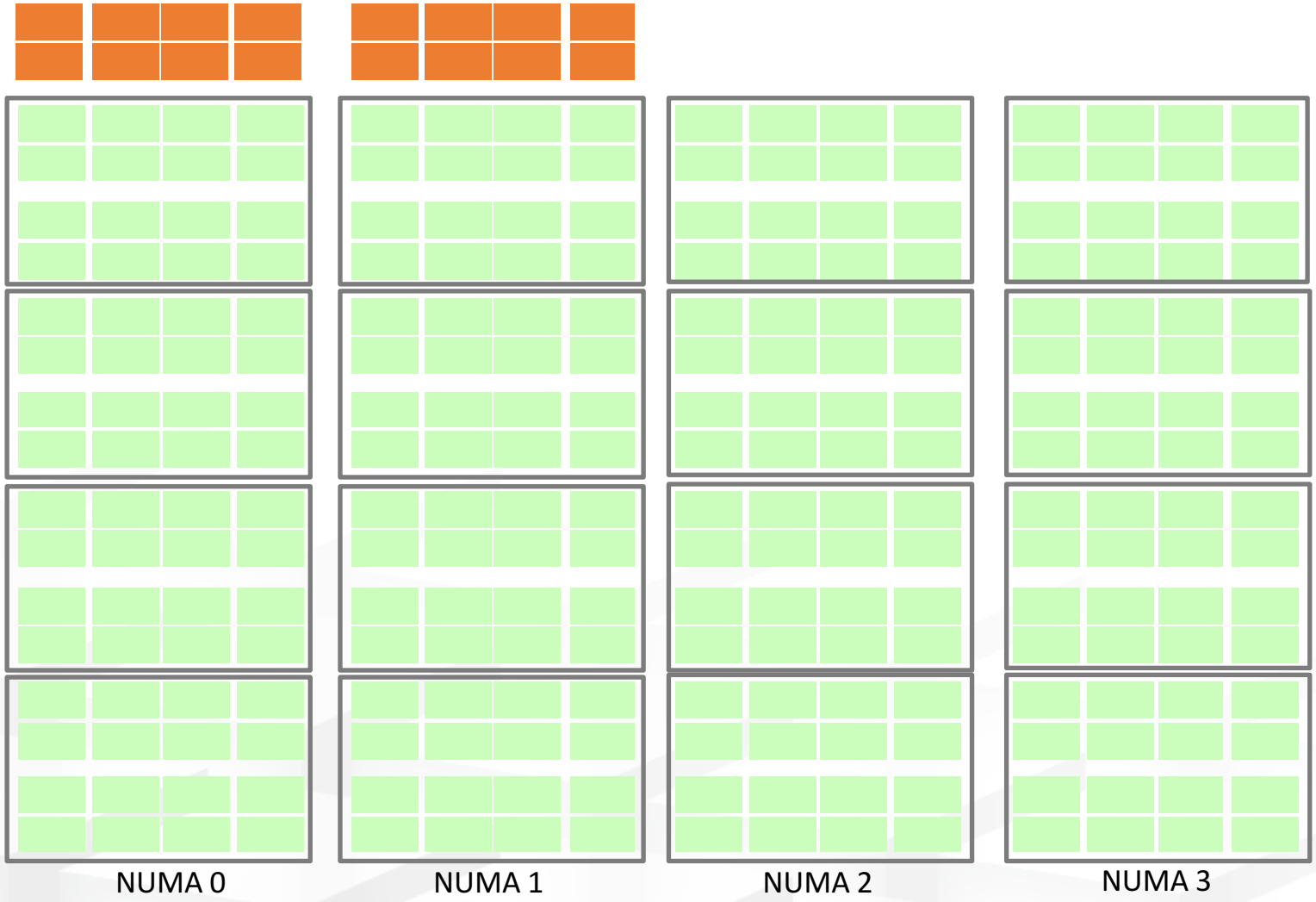
IHK/McKernel: Lightweight Multi-kernel Architecture

- **Interface for Heterogeneous Kernels (IHK):**
  - Allows *dynamic partitioning* of nod
  - Enables management of mult
  - Provides inter-kernel commu
- **McKernel:**
  - A lightweight kernel developed from scratch,
  - Designed for HPC, noiseless, simple, impl performance sensitive system calls (roughly process and memory management) and the rest ded to Linux

No Linux kernel modifications!
No node reboot during reconfiguration and LWK initialization.



Linux vs. McKernel cores on Xeon Phi KNL



- ***LWK runs on the majority of the chip***
- ***A few CPU cores are reserved for Linux***
- ***Mechanism to map inter-core communication to MPI process layout***

Linux CPU

McKernel CPU

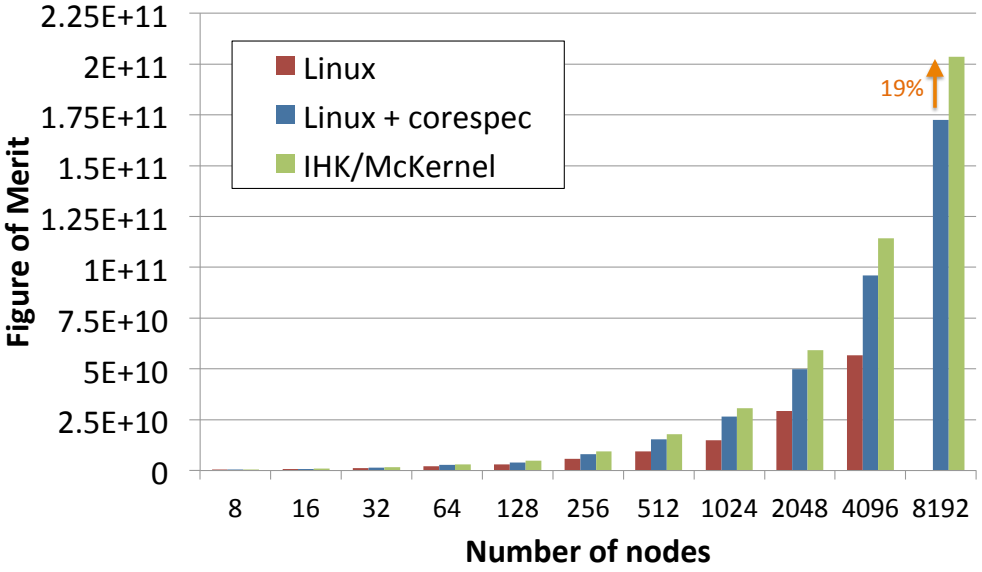
RIKEN

Oakforest-PACS Configuration

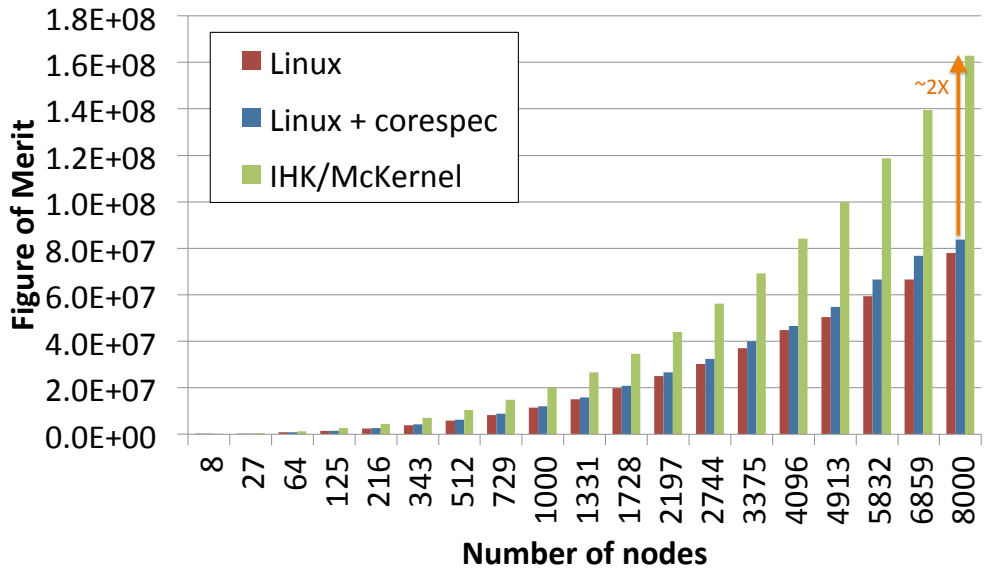
- **8k Intel Xeon Phi (Knights Landing) compute nodes**
  - Intel OmniPath v1 interconnect
  - Peak performance: ~25 PF
- **Intel Xeon Phi CPU 7250 model:**
  - 68 CPU cores @ 1.40GHz
  - 4 HW thread / core
    - 272 logical OS CPUs altogether
  - 64 CPU cores used for McKernel, 4 for Linux
  - 16 GB MCDRAM high-bandwidth memory
    - Hot-pluggable in BIOS
  - 96 GB DRAM
  - **Quadrant flat mode**

RIKEN

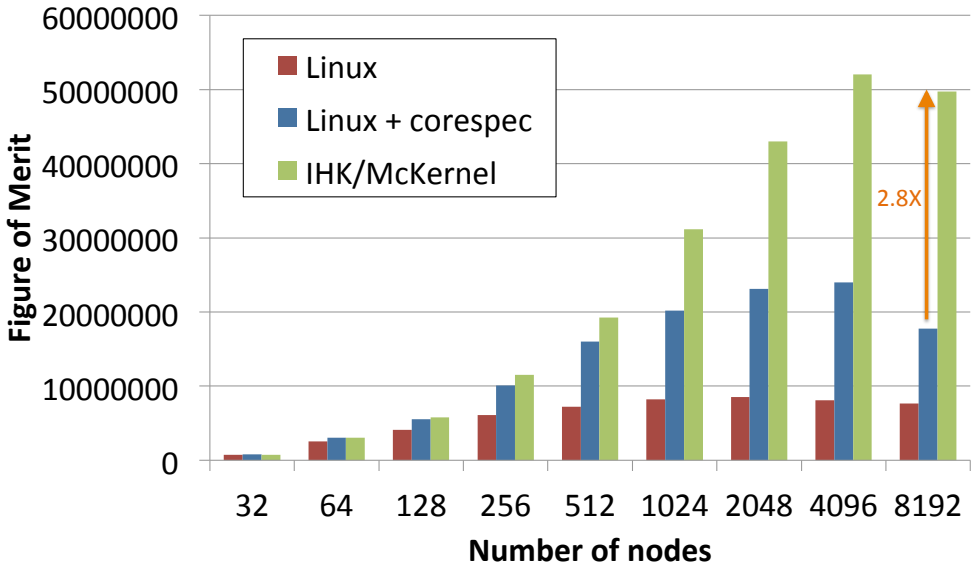
Mini-applications on full-scale OFP



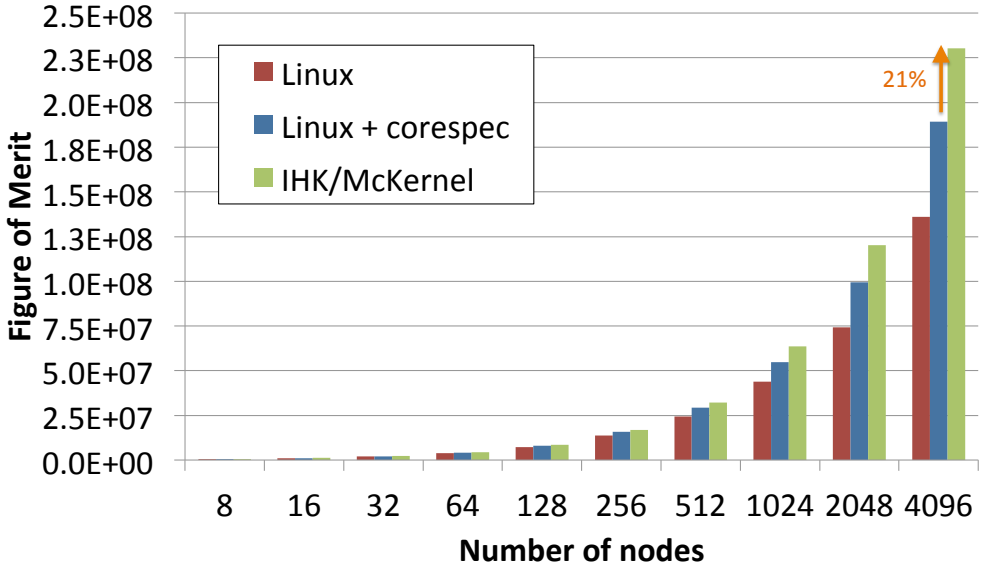
AMG2013



Lulesh



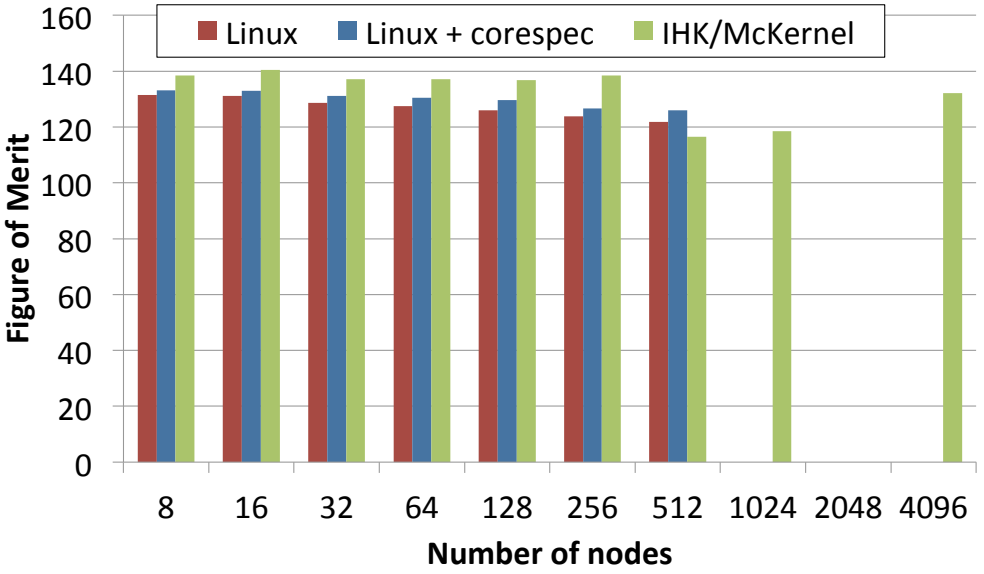
MiniFE



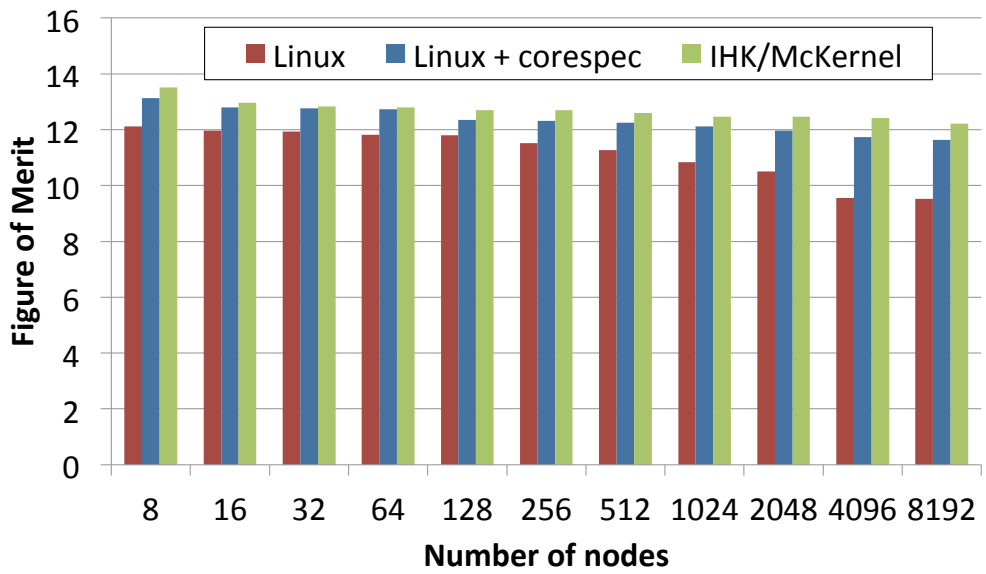
MILC

RIKEN

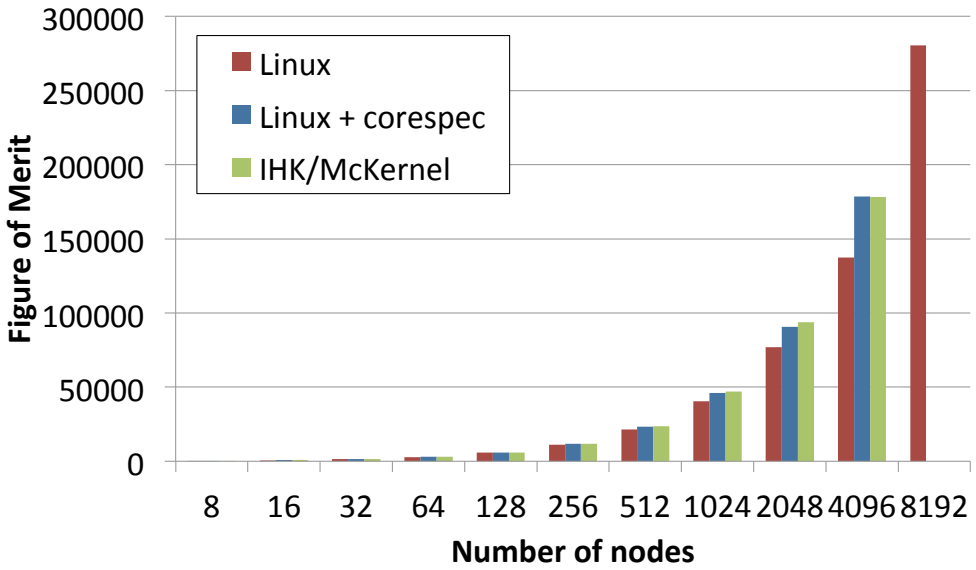
Mini-applications on full-scale OFP



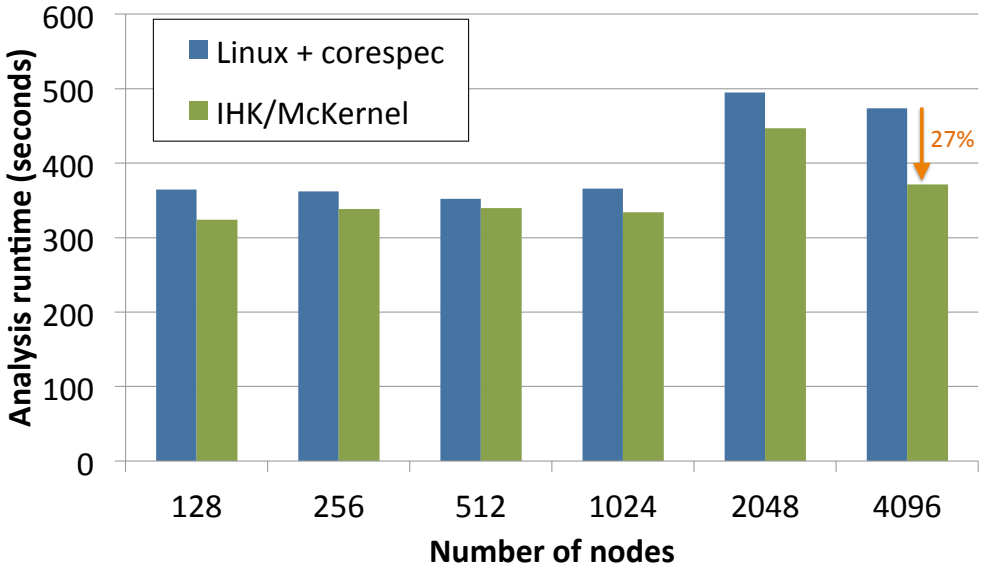
LAMMPS



GeoFEM



HPCG



GAMERA

Thank you for your attention!
Questions?