TECHNISCHE UNIVERSITÄT DRESDEN

ZIB

The Hebrew University of Jerusalem

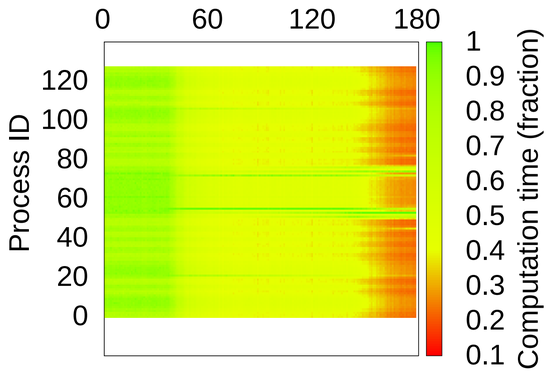
FFMK: A FAST AND FAULT-TOLERANT MICROKERNEL-BASED SYSTEM FOR EXASCALE COMPUTING

Phase 1 (2013 – 2015)

Phase 1 Results: Summary

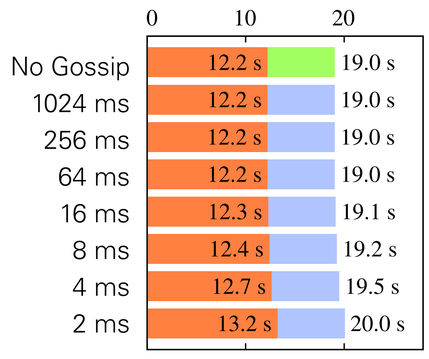
- First L4-based prototype
- Several source-compatible MPI applications ported
- Tested on small island of real HPC cluster
- Gossip scalability and resilience modeled, simulated, and measured
- Erasure-coded in-memory checkpoints with XtreemFS, tested on Cray XC40
- 2 SPPEXA Workshops

Load Imbalances



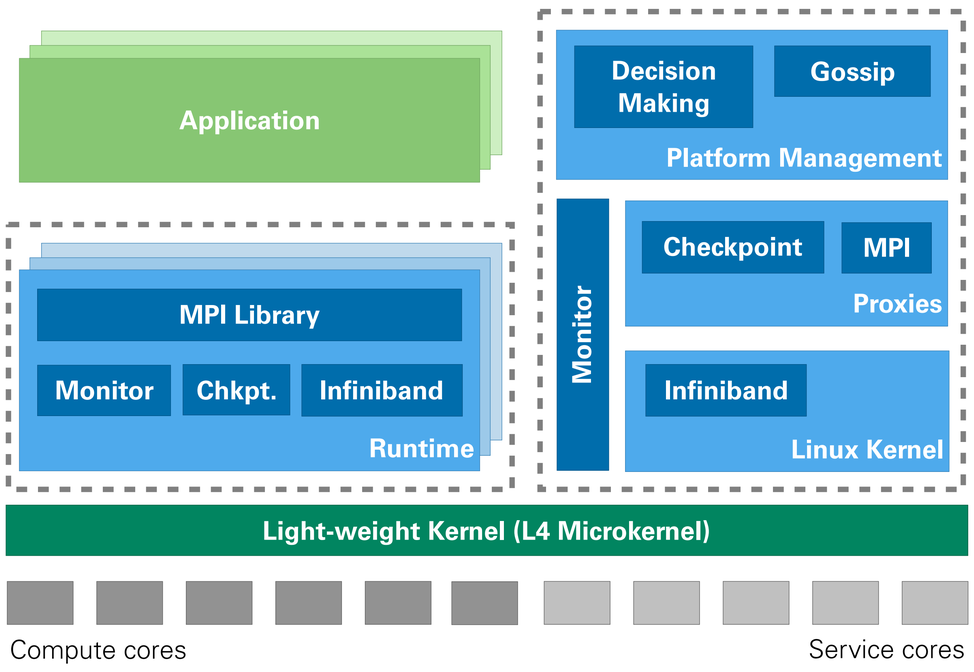
COSMO-SPECS+FD4 (128 ranks,180 timesteps)

Gossip: Scalability/Overhead



MPI-FFT (Blue Gene/Q, 1024 nodes)

Phase 2 (2016 – 2017)

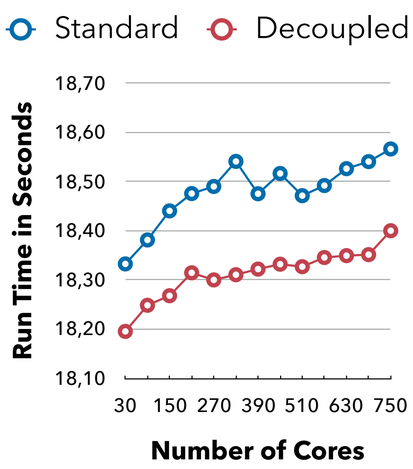


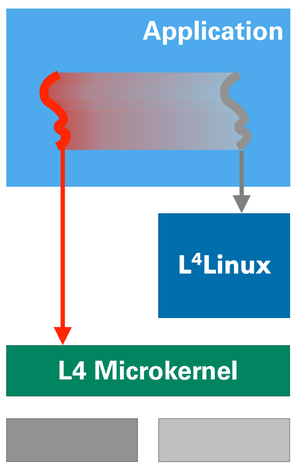
FFMK System Architecture

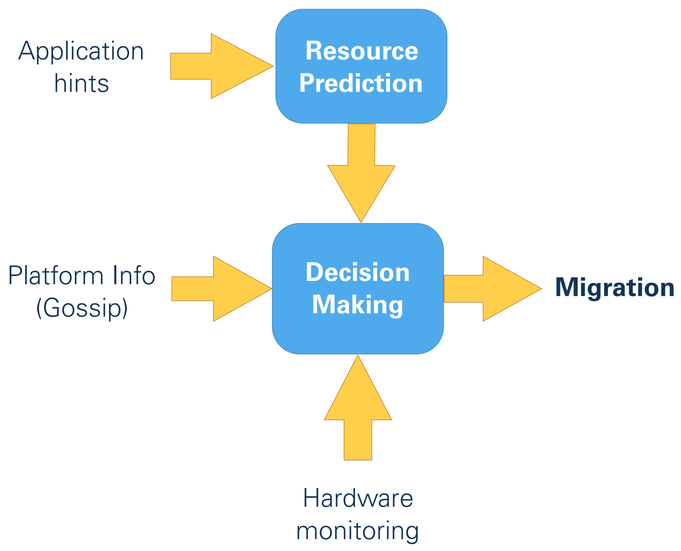
- L4 microkernel on every node
- Programming paradigms provided as library-based runtimes
- Performance-critical parts of MPI, InfiniBand, and checkpointing run directly on L4
- Non-critical support functionality reuses Linux (e.g., XtreemFS MRC+OSD, MPI startup+control)
- Gossip algorithms disseminate info for platform management
- Linux compatibility via virtualization
- Optional application hints can improve decision making
- GROMEX, COSMO-SPECS+FD4, CP2K, benchmarks, mini apps, …

Second L4-based Prototype: Decoupled Execution

- Avoids operating system noise by sidestepping Linux
- HPC Applications are ordinary Linux processes, but its threads moved to compute core controlled by L4
- Communication via InfiniBand via direct hardware access
- Linux System calls: Move thread back into Linux, handle operation on service core, then return to compute core
- L4 system calls: faster scheduling, threads, memory, …







Dynamic Platform Management

- Consider CPU cycles, memory bandwidth, and other resources
- Classification based on memory load („memory dwarfs“) to optimize scheduling and placement
- Prediction of resource usage using hardware counters and application-level hints (e.g., number of particles, time steps)

Fault Tolerance

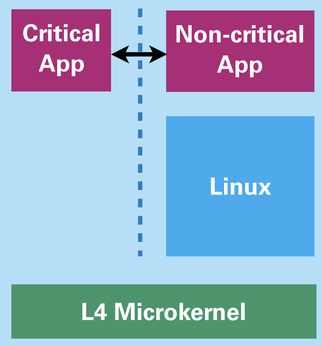
- Application interfaces to optimize or avoid C/R (e.g., hints on when to checkpoint, ability to recover from node loss)
- Node-level fault tolerance: Multiple Linux instances, micro-rebooting, proactive migration away from failing nodes

Prof. Hermann Härtig
Technische Universität Dresden

Carsten Weinhold
Adam Lackorzynski
Jan Bierbaum
Martin Küttler
Maksym Planeta
Hannes Weisbach

Operating Systems

- The Performance of µ-Kernel-Based Systems, SOSP 1997
- VPFS: Building a Virtual Private File System with a Small Trusted Computing Base, EuroSys 2008
- ATLAS: Look-Ahead Scheduling Using Workload Metrics, RTAS 2013

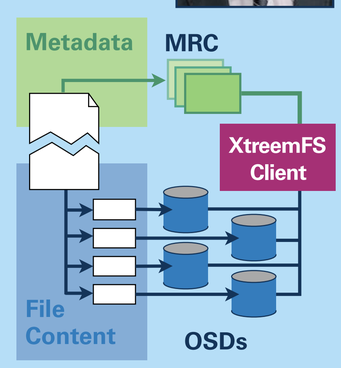


Prof. Alexander Reinefeld
Zuse Institute Berlin

Thomas Steinke
Thorsten Schuett
Florian Wende

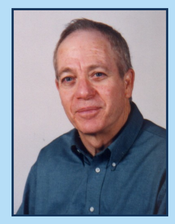
Distributed File Systems

- Flease - Lease Coordination Without a Lock Server, International Parallel and Distributed Processing Symposium, 2011
- Consistency and Fault Tolerance for Erasure-Coded Distributed Storage Systems, Workshop on Data Intensive Distributed Computing at HPDC 2012



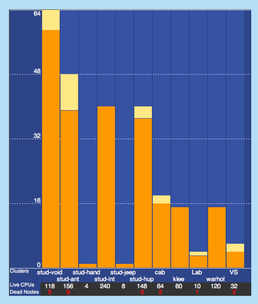
Prof. Amnon Barak
Hebrew University of Jerusalem

Amnon Shiloh
Ely Levy
Tal Ben-Nun
Alexander Margolin
Michael Sutton



Load Balancing

- Resilient Gossip Algorithms for Collecting Online Management Information in Exascale Clusters, Concurrency and Computation: Practice and Experience, 2015
- An Opportunity Cost Approach for Job Assignment in a Scalable Computing Cluster, IEEE Transactions on Parallel and Distributed Systems Vol. 11, 2000

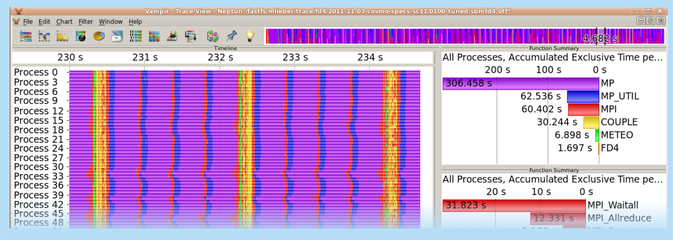


Prof. Wolfgang E. Nagel
Technische Universität Dresden, ZIH

Matthias Lieber

MPI and Performance Analysis

- The International Exascale Software Roadmap, International Journal of High Performance Computer Applications 25(1), 2011
- VAMPIR: Visualization and Analysis of MPI Resources, Supercomputer 63, XII(1):69–80, 1996



Scientific Network (Selection)

Center for Advancing Electronics Dresden
TU Dresden Excellence Cluster

Highly Adaptive Energy-Efficient Computing
SFB912

ASTEROID
SPP1500

Frank Bellosa
Karlsruhe Institute of Technology

Gernot Heiser
UNSW, NICTA

Michael Bussmann
Helmholtz Zentrum Dresden Rossendorf

Laxmikant V. Kale
University of Illinois at Urbana-Champain, Charm++

Vijay Saraswat
IBM Research Zürich, X10

Eric Van Hensbergen
IBM Research Austin
DARPA HPCS, FastOS, X-Stack

Yutaka Ishikawa
University of Tokyo
RIKEN

Torsten Hoefler
ETH Zurich

Frank Mueller
North Carolina State University

Related Projects

Argo / Hobbes / mOS
Argonne / Sandia / Intel

SPPEXA: ESSEX / GROMEX
Gerhard Wellein / Ivo Kabadshow

SPPEXA