

Kerrighed: A SSI Cluster OS Running OpenMP



EWOMP 2003

David Margery, Geoffroy Vallée, Renaud Lottiaux,
Christine Morin, Jean-Yves Berthou
IRISA/INRIA – PARIS project-team

EDF R&D



**Research
& Development**

Introduction



- OpenMP only available on SMP machines in industry
- Ongoing research on running OpenMP programs on clusters
 - Functionality
 - Kerrighed
 - Performance

Kerrighed Approach



- Single system image operating system for clusters
 - Virtual SMP
 - Simple to get an OpenMP program up and running
 - Validation of Kerrighed SSI properties
- Tool to tune parallel programs for clusters
 - Identification of performance bottlenecks
 - Profiling or pedagogical use
- Various OpenMP compilers can be used
 - Kerrighed is independent of the programming environments
 - Sound basis to experiment cluster-aware compilers

Kerrighed Overview



- Kerrighed provides a full Posix thread interface on a Linux cluster
 - Combining ease of use and high performance on small clusters (up to 128 nodes)
 - Kernel level implementation of a set of distributed services
- Global resource management
 - Containers for memory page access and sharing cluster wide
 - Global process management
 - Process and thread duplication, migration and checkpointing
 - Configurable, modular global scheduler
- All distributed services rely on a high performance portable communication system providing a kernel level interface

OpenMP on Kerrighed

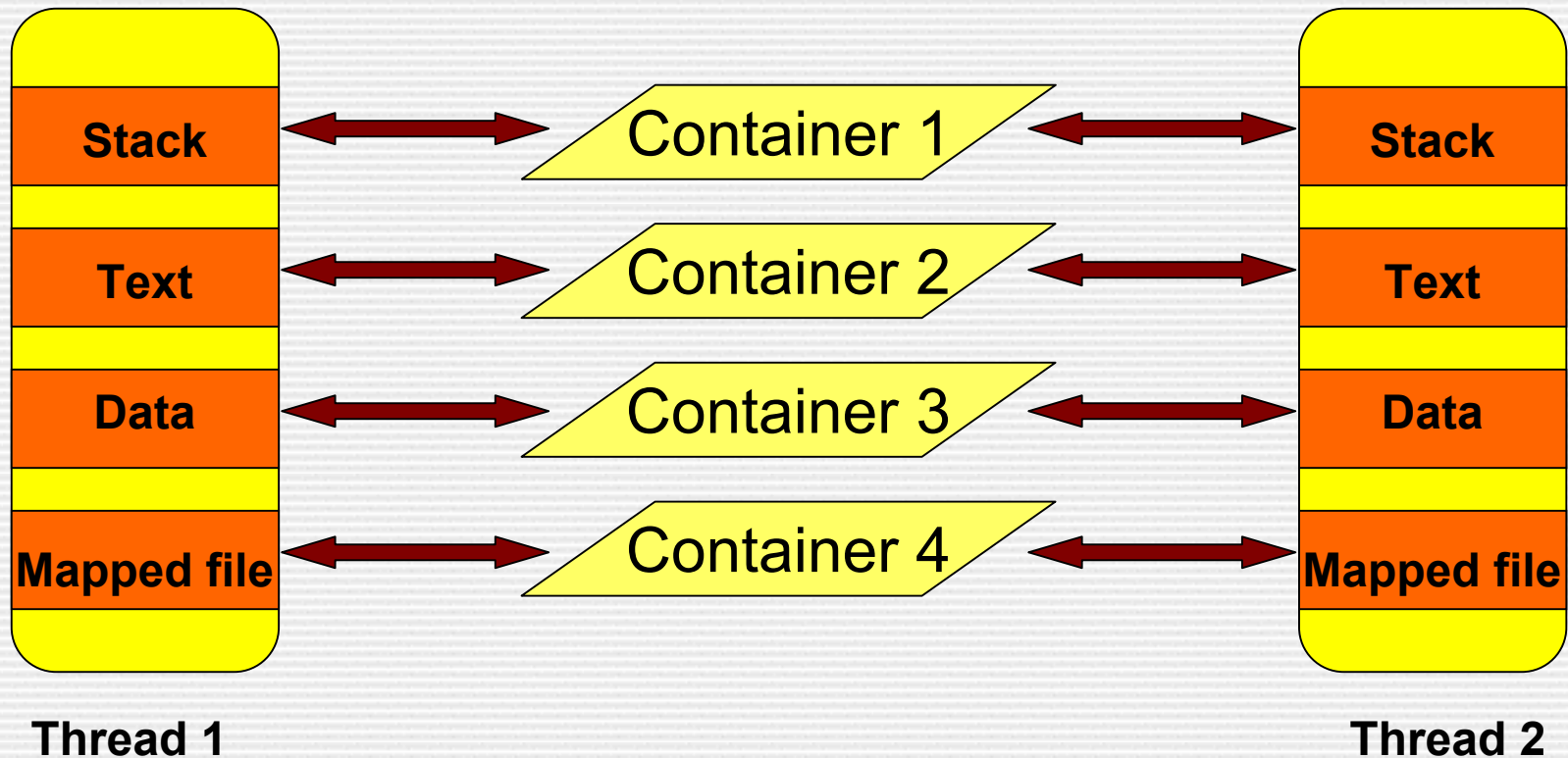


- Use of a compiler targeting Posix thread
 - The applications and the compiler are not modified
 - Standard *Pthread* library replaced by *krgthread* library proving the Posix interface on Kerrighed clusters
- Kerrighed support for OpenMP programs execution
 - Transparent thread deployment on cluster nodes
 - A thread is a Linux process
 - Containers to support shared variables in the cluster
 - Synchronization of threads executing on different nodes

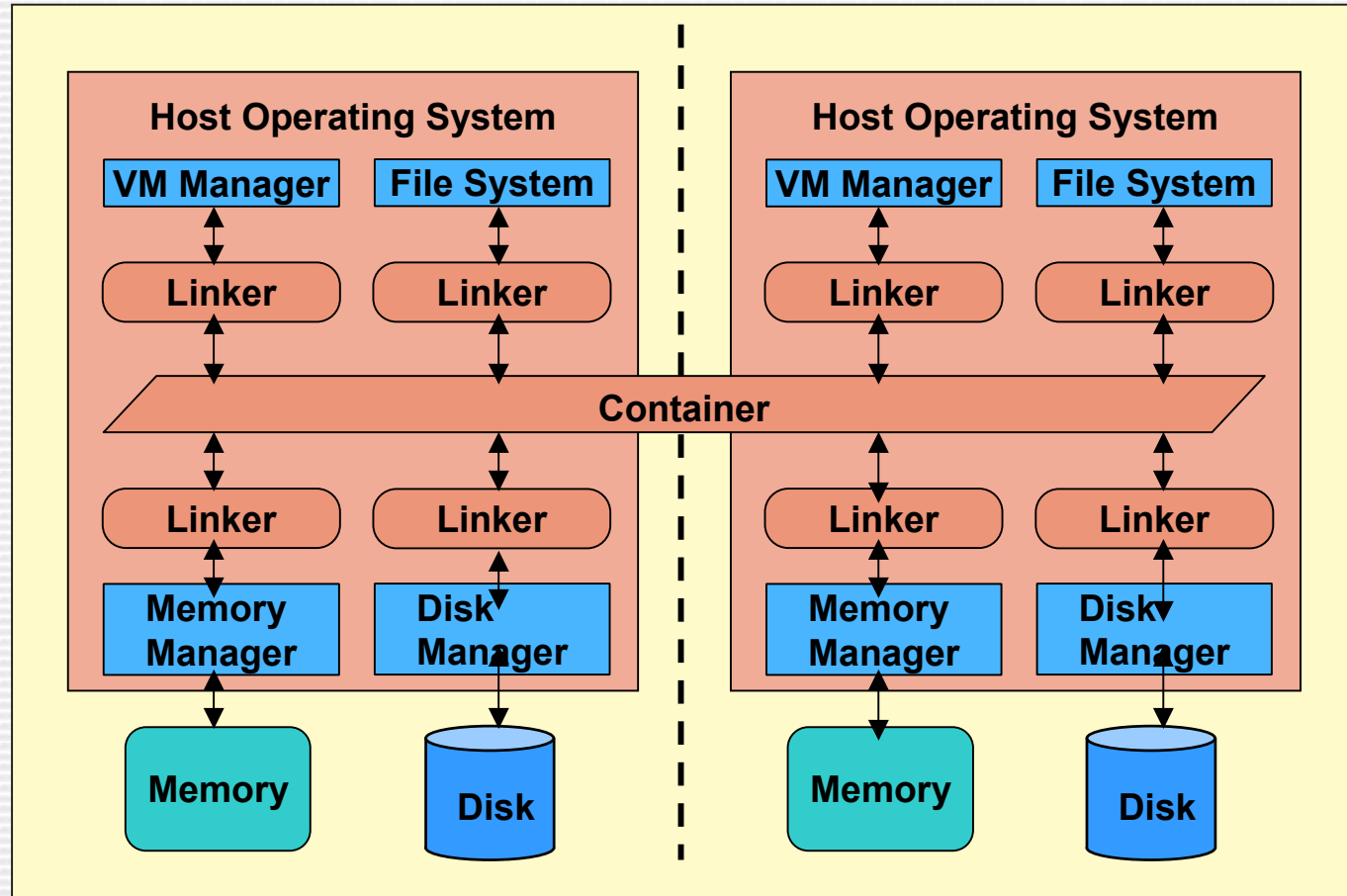
Shared Variables



- Kerrighed implements a kernel level DSM based on containers
 - Sequential consistency, page granularity
- The complete address space of a process is shared including the stack of each of its threads



DSM Implementation in Kerrighed



Private Variables



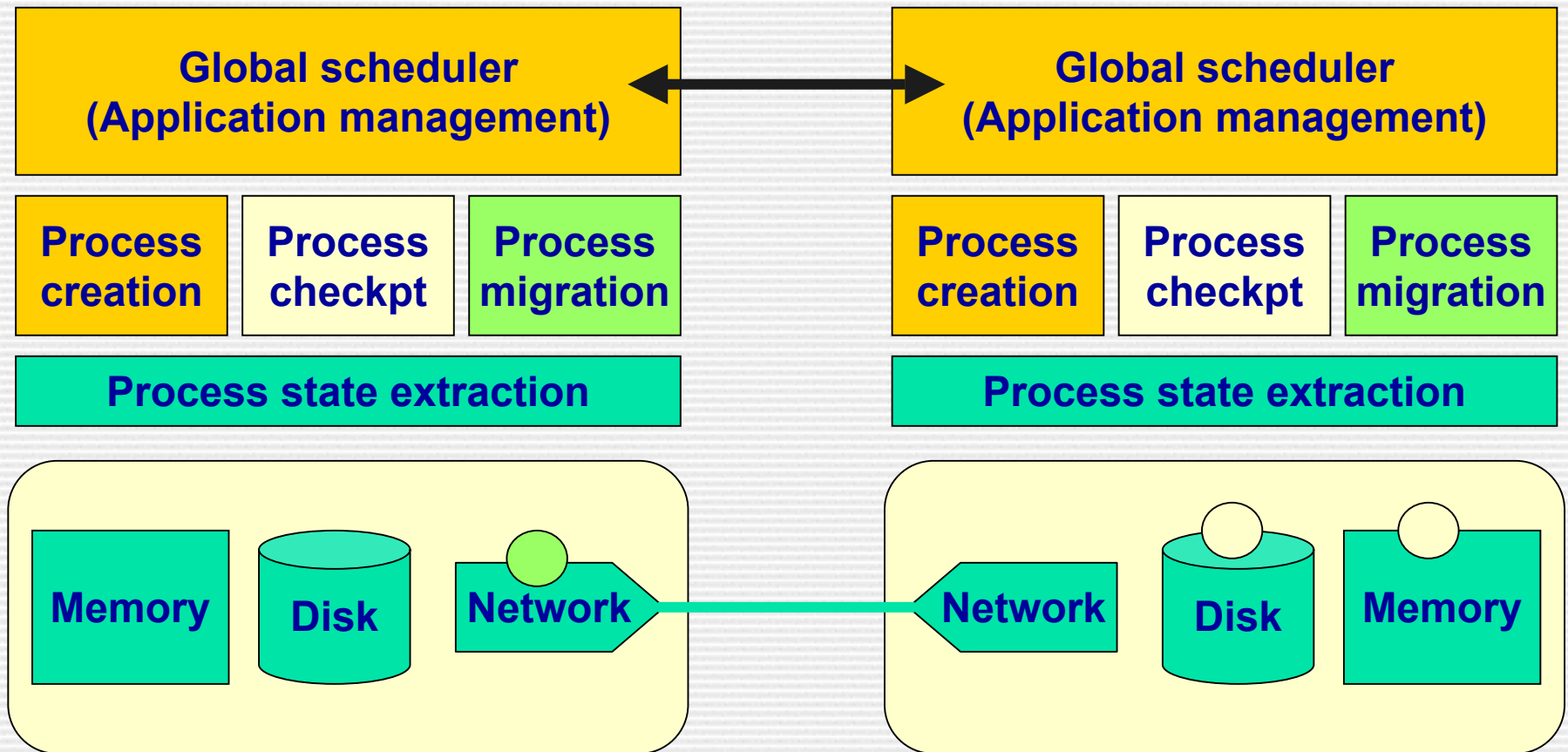
- Two options
 - Detection and redeclaration of private variables by the compiler
 - Risk of false sharing
 - Implementation based on the *thread_private_data* mechanism offered by Posix threads
 - Efficiency
- Extension of the *mmap* system call to support local memory
 - MAP_LOCAL flag
 - Allocation of the memory in the virtual address space of each thread of a process without linking it to a container
 - The same virtual address designate a different memory region on each thread

Synchronization



- Direct support of all Posix thread synchronization primitives
- Kerrighed provides natively
 - Distributed locks
 - Semaphores
 - Wait conditions
 - Barriers (Kerrighed extension to Posix thread)
- A thread using Kerrighed synchronization primitives can migrate and be checkpointed at any time
 - Except when blocked in the OS

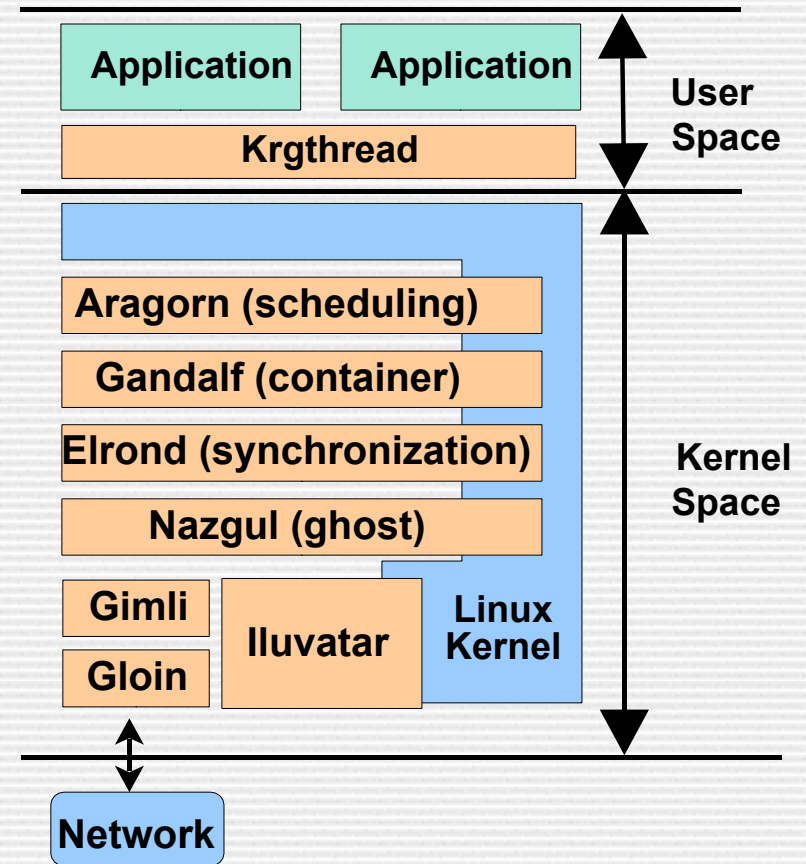
Thread Management Mechanisms



Kerrighed Implementation



- **Extension of Linux kernel**
- 7 modules
 - Process & load balancing (**Aragorn**, 13000 lines)
 - Containers (**Gandalf**, 8000 lines)
 - Synchronization (**Elrond**, 4000 lines)
 - Ghosts (**Nazgul**, 1000 lines)
 - Communication (**Gimli**, **Gloin**, 9000 lines)
 - Tools (**Iluvatar**, 2000 lines)
- Limited patch to the kernel (300 lines)
- 140 Man/Month since 1999



Experimentation



- Goal: **correctness not performance**
- Execution of OpenMP programs compiled with Omni 1.4 and linked to *krgthread* library on a 6 node cluster
- Successful execution of the 288 tests provided with Omni, of NAS benchmarks and of industrial applications (HRM1D = 7000 lines of Fortran)
 - ... but we observe a slowdown
 - Not surprising, efforts concentrated on functionalities
- **Functionalities are achieved**
- **Work on performance is a next step**

Conclusion



- Why executing OpenMP applications on Kerrighed, a SSI OS for clusters ?
 - **Easy** – Full support of Posix thread interface, no modification to the compiler or to the application
 - **Efficient global resource management** – remote paging, cooperative file cache
 - **Concurrent execution with other applications**
 - **Load balancing** – scheduling policy configurable
 - **Checkpointing** – transparent to the programmer
- Kerrighed V0.70 is available as an open source software under the GPL licence (<http://www.kerrighed.org>)
 - Kerrighed V0.70 supports the execution of unmodified OpenMP applications compiled with Omni 1.4
 - Load balancing

Future Work



- Better compilers for clusters
 - Page-aware compilers
- Tools for tuning parallel algorithms
 - Understanding performance bottlenecks
- Optimization of Kerrighed
 - Memory allocation
- Kerrighed prototype
 - New functionalities (migration of sockets, PFS/DFS, efficient checkpointing, HA)
 - Linux 2.4.x, 2.5.x
 - Support of 64 bit processors



Research
& Development



<http://www.kerrighed.org>

Kerrighed is registered as a community trademark.

