



University of Karlsruhe

Clusterfile parallel file system

Florin Isaila
Guido Malpohl
Vlad Olaru
Prof. Walter F. Tichy



University of Karlsruhe

Clusterfile parallel filesystem

- **Overview/Problem Definition**
 - **ROLE: Parallel access to storage resources within a cluster**
 - **Single system image**
 - **Logical parallelism: how a file is accessed by several processors**
 - **Physical parallelism: how a file is de-clustered over several nodes**
 - **Main performance penalty: poor match between logical and physical parallelism**



University of Karlsruhe

Clusterfile parallel filesystem

- **Technical Approach**

- Files are striped over several I/O nodes
- Applications run on compute nodes
- Flexible physical distribution: arbitrary file distribution over several cluster disks

Advantage: logical parallelism matches the physical parallelism => increased performance and scalability

- Views: application-defined logical windows to a data subset of a parallel file

Advantages: accessing non-contiguous regions of a file with a single call, simplified offset computation

- Collective I/O: merge several I/O requests from different compute nodes before sending them to disks

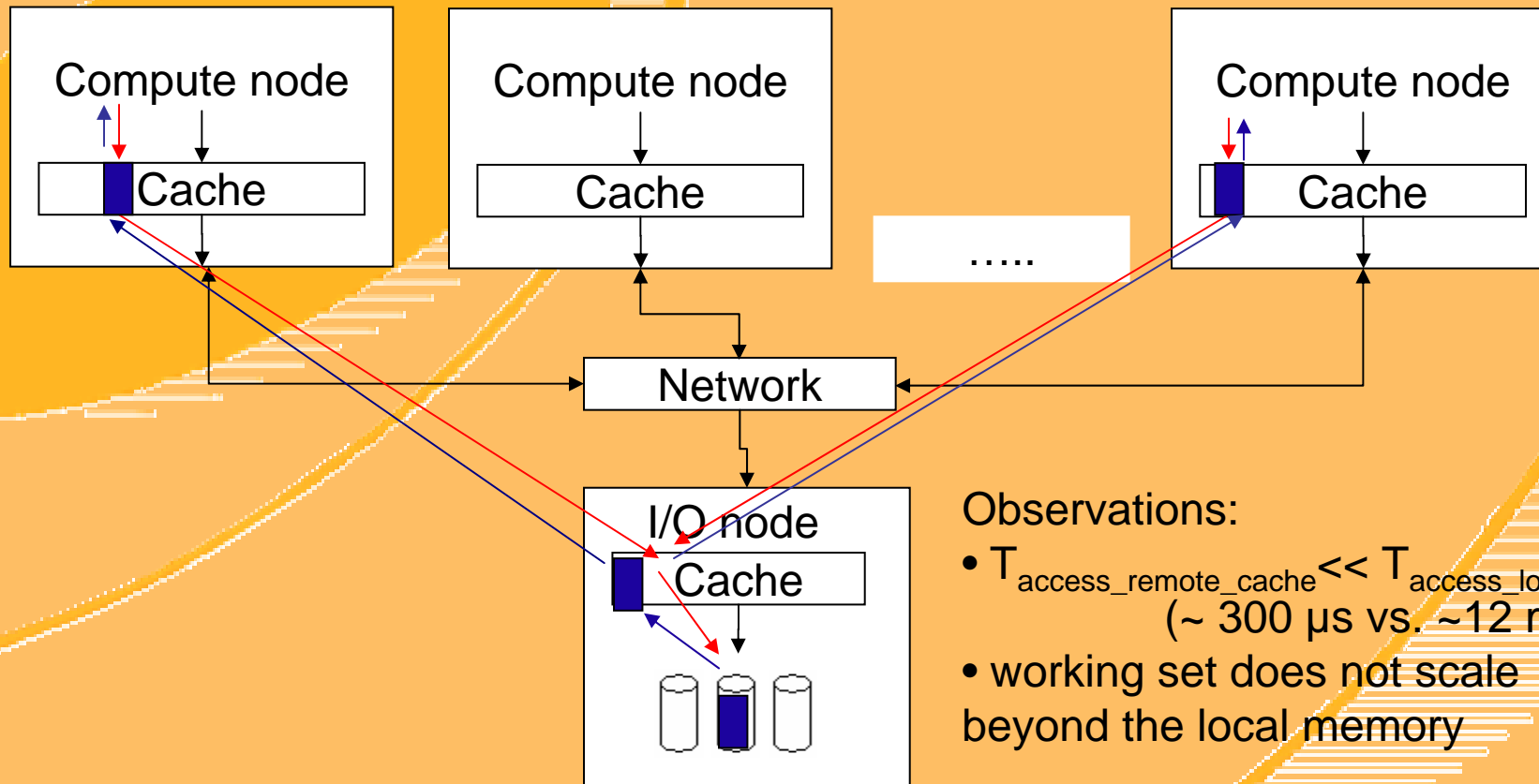
Advantage: improved network and disk throughput



University of Karlsruhe

Clusterfile parallel filesystem

- **Cooperative caching:** Joint management of distributed caches



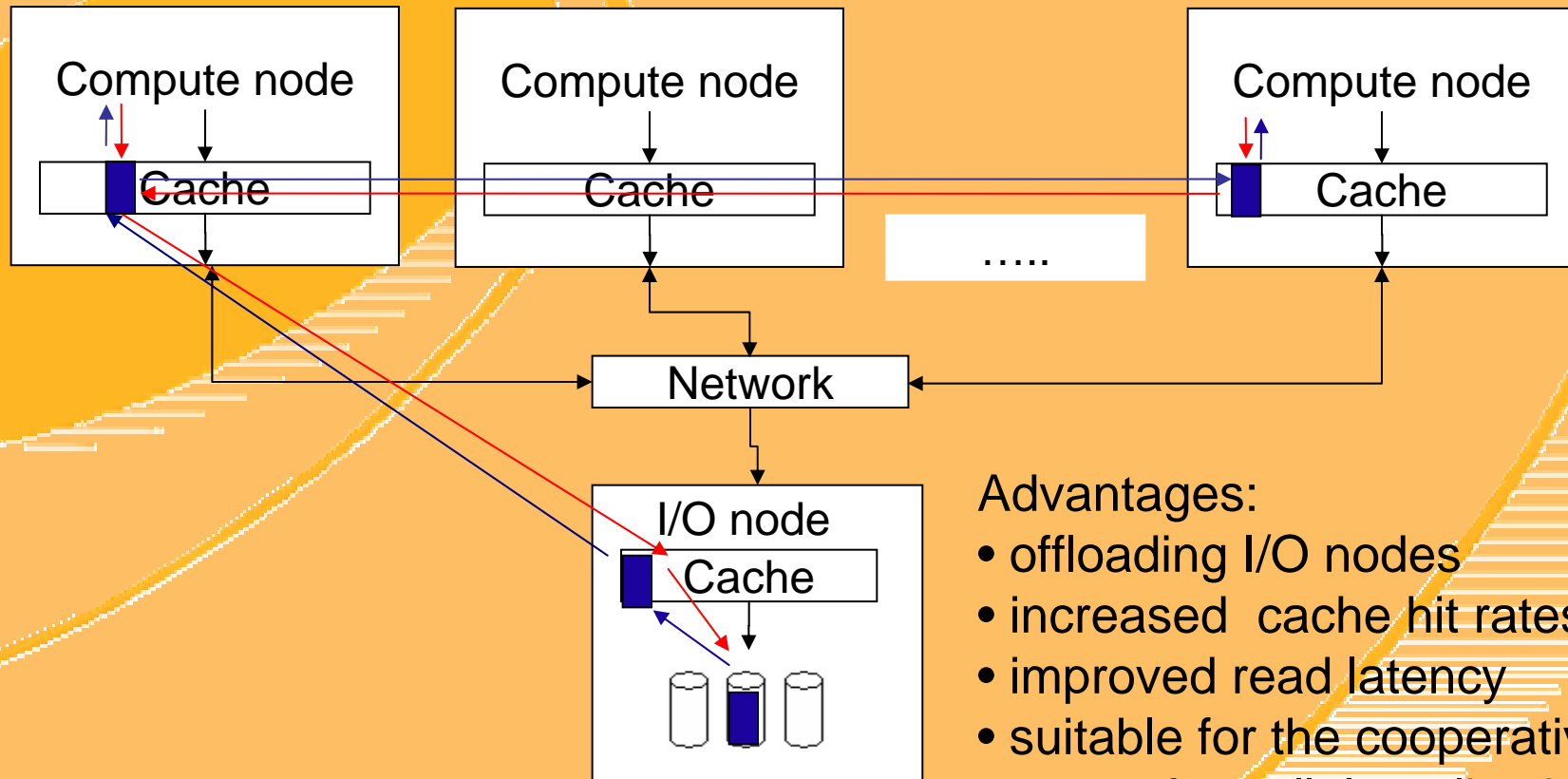
Observations:

- $T_{\text{access_remote_cache}} \ll T_{\text{access_local_disk}}$
(~ 300 μs vs. ~12 ms)
- working set does not scale beyond the local memory

University of Karlsruhe

Clusterfile parallel filesystem

- **Cooperative caching**



Advantages:

- offloading I/O nodes
- increased cache hit rates
- improved read latency
- suitable for the cooperative nature of parallel applications



University of Karlsruhe

Clusterfile parallel filesystem

- **Results**

- user-level library
- LINUX kernel interface
- MPI-IO library

- **Looking Ahead**

- metadata distribution and replication
- moving functionality to LINUX kernel
- integrate cooperative caching policies
- evaluate on large Itanium clusters