Software Approaches for Resilience of High Performance Computing Systems: A Survey

Jie JIA, Yi LIU, Guozhen ZHANG, Yulin GAO, Depei QIAN

Frontiers of Computer Science, DOI: 10.1007/s11704-022-2096-3

Problems & Ideas

- Problems:
  - System resilience has been regarded as one of the critical challenges for large-scale HPC systems.
  - Various software techniques to improve the system resilience have been proposed in recent years. However, a comprehensive survey of the state-of-the-art in software resilience is lacking.
- Ideas: An introduction and analysis of the current research in various areas of software resilience technology, as well as the challenges faced in this area.

Abnormal states of HPC systems

Class	Meaning	Typical Examples	Explanation
Fail-stop failure	Hardware and/or software stop working	Kernel Panic	Kernel error from which the operating system cannot quickly recover.
		Node Heartbeat Fault	Exception when accepting the heartbeat from other nodes.
		Traps	Segmentation Faults, Trap invalid opcode.
		GFS Failure	Failure of the global file system.
		Scheduler	Internal bugs of job scheduler.
		Acc Failure	Failure of accelerators or co-processors.
		Storage Failure	Storage system fails to work.
		Node hardware failure	Node fails due to power/cooling-system error, damage of hardware components, etc.
		Interconnect Conjunction	Network connection is congested.
Soft Error / Fail-continue error	System still works but the execution of application incorrect	SDC	Undetected silent data corruption.
		CFE	Control flow error.
		MCE	Memory check exception.

Main Contributions

- Contributions:
  - A comprehensive and systematic survey of existing software resilience approaches for HPC systems.
  - A discussion of challenges for software resilience approaches regarding recent developments of HPC systems, mainly in scalability and heterogeneous architecture.

Classification of typical resilience approaches

Resilience Method	Checkpointing	Replication	Soft Error Resilience	ABFT	Fault Detection and Prediction
Redundancy data	System memory or application data space	Process data and message	N/A	Checksum of algorithm	N/A
Recovery method	Failure-rollback	Forward recovery	Error-restart	Error-restart	N/A
Overhead/Cost	Medium	High	Medium	Low	Low
Generality	Systems and applications	Systems and applications	Systems and applications	Applications	Systems and applications
Ease of use or deployment	Easy	Easy	Hard	Hard	Medium
Limitation	Scalability	Resource consumption and scalability	Soft error only	Algorithm-dependent	Rely on other recovery methods