Diss. ETH No. 16318

Computer-Aided Management of Commodity Parts-Based Supercomputers

A dissertation submitted to the
Swiss Federal Institute of Technology Zurich

For the degree of
Doctor of Technical Sciences

Presented by Josef Nemecek
dipl. Informatik-Ingenieur ETH
Born March 31, 1973
Citizen of Wädenswil (Switzerland)
and Czech Republic

Accepted on the recommendation of
Prof. Dr. Bernhard Plattner, examiner
Prof. Dr. Anton Gunzinger, co-examiner

2005

Abstract

Supercomputers are used to solve big problems – they are «nutcrackers» that support scientists, researchers and developers in decoding the human genome, simulating the weather and climate, creating virtual wind tunnels for planes and cars, and designing effective medicaments – the so-called «grand challenge problems». Smaller supercomputers are used for tasks that require more performance than a single computer can deliver: Web-servers, data centers and calculation systems for in-house research.

One buzzword is currently changing the supercomputing market dramatically: Commodity. More and more systems are based on «commodity off-the-shelf parts»: CPUs, memories, storage devices, networking technologies and software that are used in ordinary workstations and servers are also used in these «commodity supercomputers». In the most extreme cases, the supercomputer basically consists of commercially available computers – specialized custom technologies are only used where absolutely necessary. These «clusters of workstations» (or «superclusters») are very likely to replace the conventional supercomputers installed in supercomputing centers, because many of these centers are actively investigating their potential and supporting the research in «commodity supercomputing».

However, one issue inhibits replacing supercomputers with superclusters today: The lack of a comprehensive integrated system management. Manual management using shell commands, scripts and independent tools only works for small systems with a few compute nodes. But for superclusters that consist of some thousand or million different components, computer-aided system management is mandatory. To provide high supercluster availability, this management must be fast, scalable, reliable, and secure.

This dissertation is the first complete research in this area that presents a concept for integrated and comprehensive system management of superclusters. This concept understands management as a *lifecycle* with the three phases: Design and simulation, installation, and operation – although only the operational phase is analyzed in-depth.

The expectations, fears and requirements of the people working with superclusters have been analyzed and compiled into the *eight bottlenecks and central requirements* of supercluster management: Scalability, availability, management integration, reliability, security, (low) overhead, compatibility, and cost. Also the *seven elements* of supercluster management are analyzed and presented: Control, configuration, monitoring, fault detection, trap handling, accounting, and planning. With these two tools in hand (bottlenecks and elements) it is possible to evaluate new and existing management architectures for their suitability for supercluster management.

This thesis presents some management architectures and offers a guide that allows the selection of the optimal architecture depending on system size and user requirements. One of the presented architectures has been implemented for the first Swiss-based supercluster «Swiss-T1», installed at the supercomputing center CAPA of the EPFL. This first implementation (called «COSMOS») is presented in detail in this dissertation, together with the project «Swiss-Tx» that this thesis was part of.

Of the *eight bottlenecks and central requirements*, scalability and availability have been analyzed on a theoretical level and the results have been turned into the presented management architectures – with clusters for availability and proxies for scalability. Integration and overhead are additionally covered in practice with the implementation of COSMOS. The remaining issues (reliability, security, compatibility, and cost) were taken out of the research focus and therefore neglected.

This dissertation proves that comprehensive and integrated supercluster management is an equal partner for enabling commodity supercomputing – together with networking technologies, communication libraries and distributed storage systems. The achievements of the research are blueprints of management architectures, with clustered managers for high availability, layers of proxies for scalability, tightly integrated application management for reliability, and modularity for integration. The implementation proves the concept to be correct, allowing efficient, stable, effective, and comprehensive management of the Swiss-T1 supercluster.

Zusammenfassung

Supercomputer werden gebaut, um grosse Probleme zu lösen – es sind «Nussknacker», um Wissenschaftlern, Forschern und Entwicklern bei der Entschlüsselung des menschlichen Genoms, der Simulation von Wetter und Klimaveränderungen, der Erstellung virtueller Windkanäle für Flugzeuge und Autos sowie der Entwicklung wirksamer Medikamente zu helfen – den so genannten «Grand Challenge Problems». Kleinere Systeme werden für Aufgaben verwendet, welche nicht durch einen einzelnen Computer gelöst werden können: Internet-Server, Datenserver und Rechensysteme für firmeninterne Forschung.

Der Supercomputermarkt wird momentan mit einem Schlagwort umgekrempelt: Commodity. Immer mehr Systeme basieren auf handelsüblichen Komponenten «ab der Stange»: Prozessoren, Speicher, Massenspeicher, Netzwerktechnologien und Software, welche in üblichen Workstations und Servern eingesetzt werden, werden auch auf diesen «Commodity Supercomputers» eingesetzt. Im Extremfall bestehen diese Supercomputer aus kommerziell erhältlichen Computern – spezialisierte Technologien werden nur dort eingesetzt, wo es absolut notwendig ist. Es ist sehr wahrscheinlich, dass die «Clusters of Workstations» (oder «Superclusters») die konventionellen Supercomputer ersetzen werden, welche in Supercomputing Centers installiert sind. Diese Rechenzentren betreiben aktiv Forschung im Bereich «Commodity Supercomputing», um dessen Potential abzuklären.

Supercomputer werden jedoch aus einem Grund noch nicht durch Supercluster ersetzt: Es fehlt ein umfassendes System Management. Das manuelle Management mittels Shell-Kommandi, Skripts und voneinander unabhängigen Tools funktioniert nur bei kleinen Systemen mit wenigen Rechenknoten. Für Supercluster mit einigen Tausend oder Millionen Komponenten ist jedoch ein computergestütztes Systemmanagement notwendig. Um eine hohe Verfügbarkeit des Superclusters zu erlauben, muss es schnell, skalierbar, zuverlässig und sicher sein.

Diese Dissertation ist die erste vollständige Forschungsarbeit in diesem Gebiet und präsentiert ein Konzept für ein integriertes und umfassendes Systemmanagement von Superclustern. Dieses Konzept sieht Systemmanagement als einen Lebenszyklus mit den drei Phasen Entwurf und Simulation, Installation und Verwaltung – wobei nur die Verwaltungsphase detailliert analysiert wird.

Die Erwartungen, Ängste und Anforderungen der Menschen, welche mit Superclustern arbeiten, wurden analysiert und zu den *acht Flaschenhälsen und zentrale Anforderungen* zusammengefasst: Skalierbarkeit, Verfügbarkeit, Integration der Verwaltung, Zuverlässigkeit, Sicherheit, (geringer) Zusatzaufwand, Kompatibilität und Kosten. Auch die *sie-*

ben Elemente des Supercluster Managements werden analysiert und präsentiert: Kontrolle, Konfiguration, Überwachung, Fehlererkennung, Warnsignale, Verrechnung und Planung. Mit diesen zwei Werkzeugen (Flaschenhälse und Elemente) ist es möglich, bestehende und neue Architekturen für Management Software zu evaluieren, ob sie für die Verwaltung von Superclustern geeignet sind.

Diese Dissertation präsentiert einige Management-Architekturen und bietet einen Führer an, welcher die Wahl der optimalen Architektur gestattet, abhängig von der Systemgrösse und den Anforderungen der Benutzer. Eine der präsentierten Architekturen wurde für den ersten Schweizer Supercluster «Swiss-T1» des Rechenzentrums CAPA der EPFL implementiert. Diese erste Implementation (mit dem Namen «COSMOS») wird in dieser Dissertation detailliert präsentiert, zusammen mit dem Projekt «Swiss-Tx», deren Teil diese Arbeit war.

Von den *acht Flaschenhälsen und zentralen Anforderungen* wurden die Skalierbarkeit und die Verfügbarkeit auf einem theoretischen Niveau analysiert und führen zu den präsentierten Architekturen – mit Clustern für die Verfügbarkeit und Proxies für die Skalierbarkeit. Integration und Zusatzaufwand werden in der Praxis durch die Implementation von COSMOS abgedeckt. Die verbleibenden Punkte (Zuverlässigkeit, Sicherheit, Kompatibilität und Kosten) waren nicht Bestandteil der Forschung.

Diese Dissertation beweist, dass umfassendes und integriertes Supercluster Management ein gleichberechtigter Partner ist für die Ermöglichung des «Commodity Supercomputing» – zusammen mit den Netzwerktechnologien, Kommunikationsbibliotheken und verteilten Festplattenspeichersystemen. Das Ergebnis dieser Forschung sind Baupläne für Managementarchitekturen, mit geclusterten Managern für hohe Verfügbarkeit, Schichten von Proxies für die Skalierbarkeit, straffe Integration des Applikationsmanagements für die Zuverlässigkeit sowie Modularität für die Integration. Die Implementation beweist die Korrektheit des Konzepts, welches eine effiziente, stabile, effektive und umfassende Verwaltung des Swiss-T1 superclusters ermöglichte.