Neue Features in PBS Professional

Dr. Jochen Krebs

Altair Engineering GmbH, Boeblingen, Germany

Abstract:

PBS Pro is an intelligent workload management and batch queuing solution for numerically intense compute environments. PBS Pro efficiently manages computational workload across local and distributed LINUX, UNIX, Mac and Windows based HPC environments – maximizing hardware and software utilization and job turn-around efficiency.

Many CAE users around the world rely on PBS to manage their jobs and application licenses. LS-Dyna is a common application in these environments and PBS provides tools for a close integration with this solver.

Under the current release 7.0, the functionality of PBS has been enhanced through some significant additional features such as:

- *Facility for job arrays. Job arrays are a mechanism for grouping related work and allowing the user to submit and manage these jobs as a whole*
- *Integration of MPICH on Linux 2.4 on x86/AMD64/EM64T*
- *Better support of special features on SGI Altix (CPUsets, Comprehensive System Accounting, etc.)*
- *Integration with IBM MPI on AIX 5 on POWER4 & 5 with IBM's Parallel Operating Environment (POE).*

Keywords:

Workload Management, Job Scheduling, Batch Queuing, Lastverteilung

1 Altair Engineering und PBS Professional

Das intelligente Workload Management System „PBS Professional“ wurde in den neunziger Jahren von der NASA entwickelt und von der Firma Veridian anschließend kommerziell vermarktet. Im Jahre 2003 wurde PBS (die Abk. für „Portable Batch System“) von der Firma Altair gekauft und einschließlich der Entwicklungsmannschaft übernommen. Seither hat Altair die Software in einem eigenen Geschäftsbereich „Enterprise Computing“ konsequent weiterentwickelt und kontinuierlich verbessert. Als bedeutender Anbieter von CAE Tools versteht Altair die Bedürfnisse der Anwender und kann daher eine optimale Integration von Workload Management und CAE Solvern sicherstellen.

Altair kooperiert mit führenden Hardwareherstellern und Softwarepartnern wie Dynamore, um unterschiedliche Hard- und Softwareplattformen zu unterstützen und den Kunden maßgeschneiderte Lösungen für ihre Anforderungen zu bieten.

PBS ist auch in einer „Open Source“ Variante unter dem Label „OpenPBS“ als kostenloser Download von Altair erhältlich, allerdings ist die Funktionalität von OpenPBS im Vergleich zu PBS Professional stark eingeschränkt und für einen Produktionsbetrieb daher nur bedingt nutzbar. Überdies bietet Altair technischen Support nur für die „Professional“ Variante an.

2 Grundlegende Funktionalität von PBS Professional

Die Aufgabe eines Workloadmanagement Systems ist es, die Anforderungen der Anwender nach IT-Ressourcen (ia. v.a. CPU Leistung und Applikationslizenzen) mit den real verfügbaren Möglichkeiten möglichst optimal in Einklang zu bringen, und zwar in einer Form, die dem Bedarf des Unternehmens bzw. der Abteilung auch tatsächlich gerecht wird (d.h. z. B. Jobs je nach Dringlichkeit entsprechend zu priorisieren).

Eine schematische Darstellung dieses Sachverhaltes gibt die folgende Abbildung:

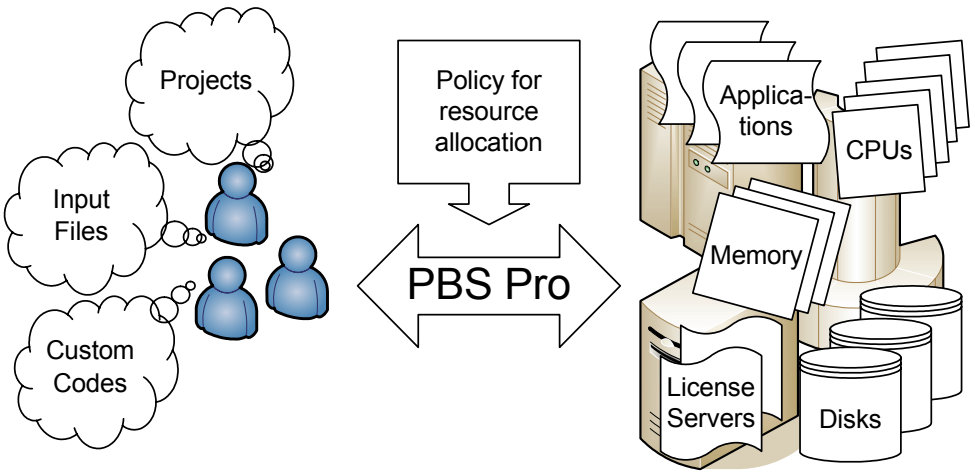

Bild 1: Aufgaben eines Workload Management Systems

Durch den Einsatz eines intelligenten Workloadmanagers kann vorhandene Hardware effizienter genutzt werden. Dies ist gerade in einem CAE Umfeld von Bedeutung, wo es typischerweise Jobs mit sehr unterschiedlichen Laufzeiten gibt. Das WMS sorgt dann dafür, dass beispielsweise kurze Testjobs mit geringen Ressourcenanforderungen bevorzugt abgearbeitet werden und sich nicht tagelang hinter langen Produktionsjobs „anstellen“ müssen.

Durch den Einsatz eines Lastverteilungssystems können auch unvorhergesehene Störungen der Hardware besser überbrückt werden. Beispielsweise werden von Hardwareausfällen betroffene Jobs automatisch auf anderen Knoten neu gestartet und der ausgefallene Knoten wird automatisch aus der verfügbaren Konfiguration entfernt, so dass der Einfluss auf den Betrieb des Gesamtsystems minimiert wird. Dies ist v.a. dann von Bedeutung, wenn der laufende Betrieb gerade nicht von einem Operator überwacht wird, also z.B. in der Nacht oder am Wochenende.

Lastverteilungssysteme sorgen über „Fairshare“ Mechanismen dafür, dass jeder Benutzer bzw. jede Benutzergruppe im zeitlichen Mittel die Ressourcenanteile zugewiesen bekommt, die seinen bzw. ihren Anforderungen entspricht. Über sog. „Preemption“ Mechanismen können jedoch auch Jobs mit höherer Priorität vorgezogen werden, z.B. wenn Deadlines für ein Projekt eingehalten werden müssen oder kurzfristig kleinere Testläufe vorgezogen werden sollen.

Aus Sicht der Benutzer erleichtert das WMS die Arbeitsabläufe durch standardisierte Skripte und stellt sicher, dass die Parameterwerte für die Solver optimal gesetzt sind. Der Zugriff auf unterschiedliche Hardwareplattformen gestaltet sich dadurch weitgehend transparent und das Risiko von Fehlern wird deutlich reduziert. Fortgeschrittene Systeme verfügen auch über die Möglichkeit, Jobs zwischen den Rechnern verschiedener Standorte auszutauschen („Peer-to-Peer Scheduling).

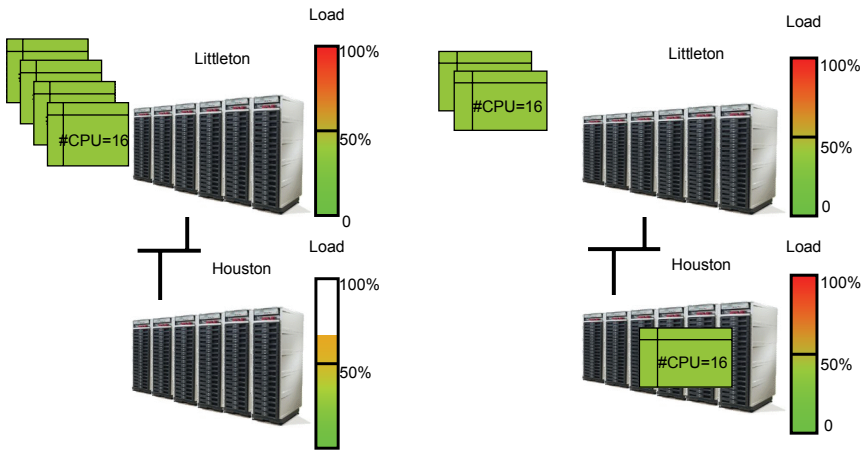

Bild 2: Peer-to-Peer Scheduling

Einen Überblick über die wichtigsten Features von PBS Professional gibt die Abb. 2.

- **Well suited for heterogeneous environments**
  - Wide range of supported architectures
- **Extended set of scheduling options**
  - Arbitrary resource awareness
  - Fair share
  - Preemptive scheduling
  - Multi-cluster scheduling
  - Job dependencies
  - Backfilling
  - Advance Reservations
- **Cycle harvesting on workstations**

- **Open interfaces to applications and other middleware**
  - MPI distributions
  - Globus, UNICORE
- **Advanced internal communication architecture**
  - Extended fault tolerance
  - Fail-over of central services
- **Integrated with (external) checkpoint and restart facilities**
- **Extended logging and accounting**
- **Security and access control lists**

Bild 3: Features von PBS Professional

Ein wesentlicher Vorteil von PBS ist das einfache Lizensierungsschema. PBS wird nach der Anzahl der genutzten CPUs lizensiert, wobei der Nutzer die Wahl zwischen Kauf- und Leasinglizenz hat. Sämtliche Funktionalität von PBS ist in dieser einen Lizenz enthalten, d.h. die entstehenden Kosten sind sehr einfach überschaubar. Die Erfahrung zeigt, dass sich die Anschaffungskosten für PBS durch die damit verbundenen Produktivitätsvorteile in der Regel nach spätestens einem Jahr amortisiert haben.

3 e-Compute – die grafische Benutzeroberfläche von PBS Professional

Gegenwärtig benutzen die meisten Anwender von PBS Professional eine kommandozeilenbasierte Schnittstelle, das sog. PBS-Submit Kommando. Es ermöglicht einen einfachen Zugriff auf alle vorhandenen Solverpakete und eine solverunabhängige Spezifikation der benötigten Resourcenparameter (z.B. Anzahl der CPUs, Scratch Space, inputdeck, etc.). PBS-Submit kümmert sich auch um den Transfer eventuell benötigter Daten- und Include Files, führt nach Jobende oder -abbruch die notwendigen Aufräumarbeiten durch und stellt die Outputfiles wieder auf dem Userterminal zur Verfügung.

Trotz dieser einfachen Bedienbarkeit kann es zweckmäßig sein, auch eine plattformunabhängige grafische Benutzeroberfläche zur Verfügung zu haben, z.B. wenn die Anwender an Windows Workstations arbeiten, ihre Jobs aber auf UNIX Systemen gerechnet haben wollen.

Für solche Zwecke bietet Altair Engineering das grafische Webportal e-Compute an.

e-Compute gestattet eine Strukturierung der typischen CAE-Arbeitsabläufe auf der Basis persönlicher Profilinformationen. Es lässt sich flexibel an individuelle Applikationen und Rechnerumgebungen anpassen und ermöglicht mittels der grafischen Benutzeroberfläche einen transparenten und plattformunabhängigen Zugriff auf unterschiedliche Rechnerplattformen. Damit wird die Fehlerrate beim Aufsetzen von Jobs weiter reduziert, außerdem können laufende Jobs mittels e-Compute besser überwacht werden. Die grafische Benutzeroberfläche verkürzt die Einarbeitungszeiten der Entwickler und erlaubt sicheren und einfachen Zugriff von inner- und außerhalb des Netzwerkes. E-Compute implementiert sicheren Datentransfer über das Internet mittels HTTP-Protokoll über SSL. Das integrierte Accounting ermöglicht eine genaue Analyse und Kostenkontrolle von Jobs und Rechnerressourcen.

Im Einzelnen stellt e-Compute den LS-Dyna Nutzern die folgenden Funktionalitäten zur Verfügung:

- *Vorgegebene Profile (entweder Nutzer- oder Site-definiert), um das Aufsetzen der Jobs zu beschleunigen*
- *Automatische Einrichtung eines Arbeitsverzeichnisses entweder lokal oder auf dem Server*
- *Transfer der notwenigen Datenfiles in das Arbeitsverzeichnis*
- *Datenfiles können zur Laufzeit des Jobs interaktiv geladen werden*
- *Senden von verschiedenen Signalen an LS-DYNA zum Beeinflussen der Jobs*
- *Anhalten, Suspendieren oder Freigeben von Jobs*
- *Anzeigen von Jobzustand und Jobdetails*

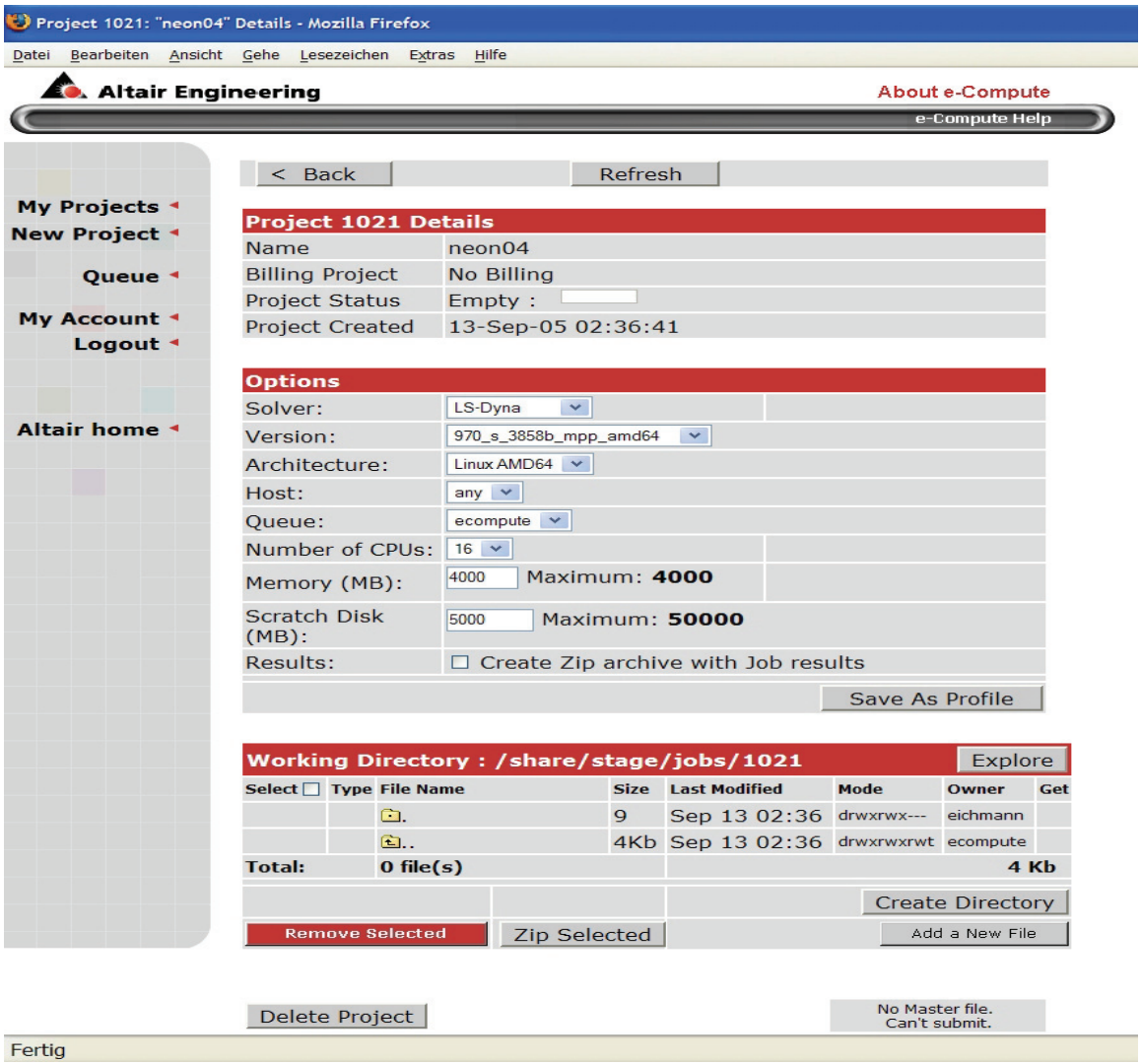

Bild 4: e-Compute Benutzeroberfläche

4 Weitere Aspekte der Integration von PBS Professional und LS-Dyna

Aus Sicht der IT-Architektur ist ein WMS in der Schicht zwischen dem Betriebssystem (heute meistens eine LINUX Variante) und den Applikationen (also z.B. den CAE-Solvern) angesiedelt und gehört damit zur sog. „Middleware“. Lastverteilungssysteme lassen sich im Prinzip unabhängig von der Applikation betreiben, allerdings kann eine engere Ankopplung für bestimmte Aufgabenstellungen sinnvoll sein.

Hier sollen v.a. drei Aspekte einer engeren Integration zwischen PBS und LS-Dyna betrachtet werden:

4.1 Lizenzverwaltung

In vielen CAE Umgebungen ist die Anzahl der verfügbaren Solver-Lizenzen limitiert. Es ist daher notwendig und sinnvoll, dass ein WMS die Verfügbarkeit der erforderlichen Lizenzen abprüft, bevor ein Job tatsächlich gestartet wird. Andernfalls könnte es zu unbeabsichtigten Jobabbrüchen kommen oder es kann passieren, dass ein Rechner mit Jobs überladen wird, weil ein Job, der keine Lizenz hat, zunächst auch keine CPU Zyklen konsumiert.

In PBS ist es möglich, beliebige benutzerdefinierte Ressourcen einzuführen, die dann von dem System entsprechend verwaltet werden. Damit ist es sehr einfach, die Verfügbarkeit von Applikationslizenzen zu überprüfen. Die Abfrage des Flexlm-Lizenzservers ist innerhalb des sog. „PBS Toolkits“ in Form eines Perl-Skriptes implementiert und geschieht transparent für den Anwender. Dieser spezifiziert in einer Kommandozeile lediglich sein Inputdeck und z.B. die Anzahl der benötigten CPUs.

4.2 Checkpointing von Jobs

PBS verfügt über die Möglichkeit, Jobs während der Verarbeitung an bestimmten Stellen („Checkpoints“) anzuhalten und zu einem späteren Zeitpunkt von dort wieder neu aufzusetzen. Voraussetzung ist, dass die fragliche Applikation ein solches Verfahren auch unterstützt, was bei LS-Dyna der Fall ist. Auch dieser Vorgang des applikationsspezifischen Checkpointing wird über entsprechende Skripte unterstützt, die bei Eintreffen eines bestimmten Signals (z.B. „Suspend Job“) eine vordefinierte Aufgabe ausführen. Das LS-Dyna Checkpointing Signal stoppt alle betroffenen Prozesse in kontrollierter Form und gibt die Applikationslizenzen zurück. Diese Lizenzen können dann für einen neuen Lauf (z.B. für einen Job mit höherer Priorität) temporär genutzt werden. Auch der Fall eines Jobabbruchs wird mittels der geeigneten lstc-Kommandos von PBS durchgeführt.

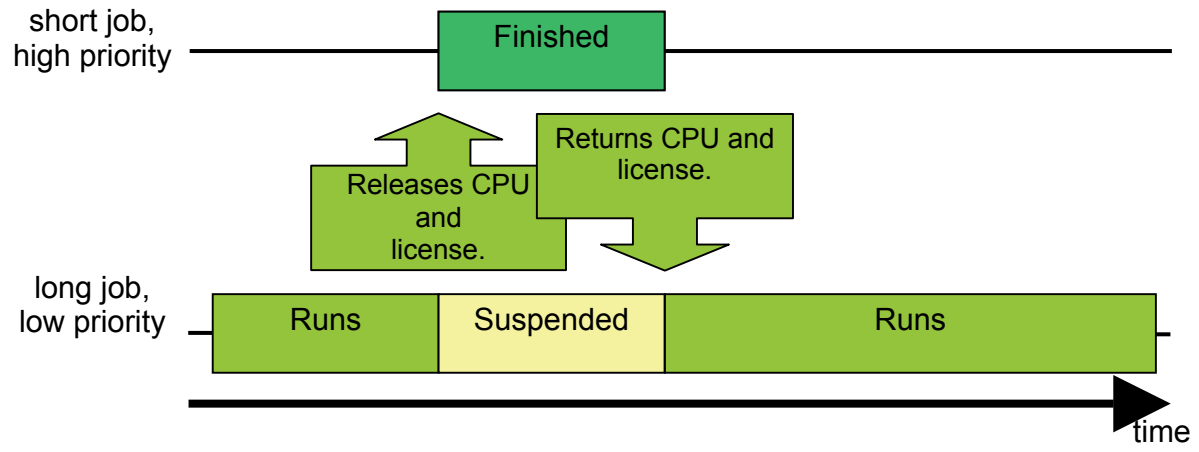

Bild 5: Job Preemption

4.3 Behandlung von Multinode Jobs

In diesem Bereich wurden mit der aktuellen PBS Professional Release 7.0 entscheidende Verbesserungen eingeführt, die im nächsten Abschnitt genauer vorgestellt werden.

5 Neue Features in PBS Professional 7.0

Eine umfassende Diskussion der neuen Features in v 7.0 würde den Rahmen dieses Papers sprengen; hier soll nur ein kurzer Überblick gegeben werden, ohne den Anspruch auf Vollständigkeit zu erheben.

5.1 Job Arrays

Ein Jobarray ist eine Zusammenfassung von Jobs, die sich nur durch einen Indexparameter unterscheiden. Ein solches Konstrukt bietet zwei Vorteile: Zum einen lässt sich ein Jobarray als Ganzes verwalten, d.h. Jobarrays können wie ein normaler Job gestartet, modifiziert und überwacht werden. Außerdem verbessern Jobarrays die Performance des Schedulings, da das Batchsystem die Metadaten eines Jobarrays effizienter verwalten kann, als dies der Fall wäre, wenn es sich um Einzeljobs handeln würde. Dies reduziert insbesondere den I/O-Aufwand für die Jobverwaltung.

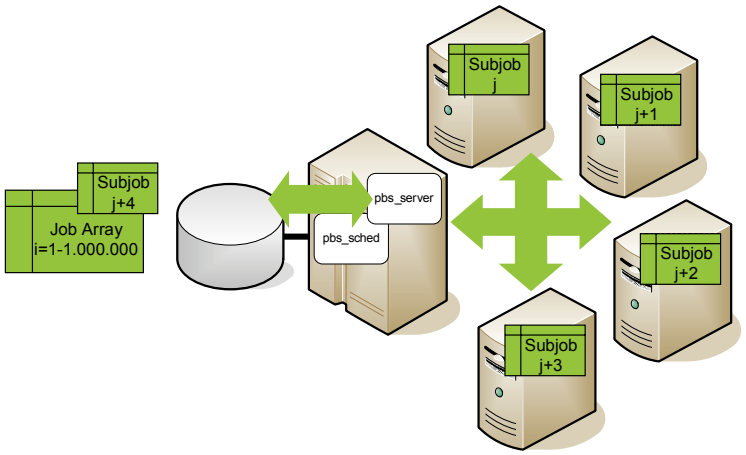

Bild 6: Job Arrays

Solche Jobarrays können genutzt werden, um Tausende von ähnlichen Jobs simultan zu starten und zu verwalten. Anwendungen gibt es im Bereich der Bildbearbeitung (Rendering), Bioinformatik oder der Kollisionsanalyse.

Innerhalb eines Jobarrays können auch die Subjobs über ihren Index einzeln angesprochen und manipuliert werden. Fairshare, Preemption, User Limits und Prolog/Epilog Skripte beziehen sich auf die einzelnen Subjobs.

5.2 Neue Features für Multinode Jobs

PBS gestattet die Allokation individueller Rechenknoten exklusiv für einen bestimmten Job („exclusive access“ oder „space sharing“). Fordert ein Job exklusiven Zugriff, wird der gesamte Knoten (unabhängig von Anzahl der Prozessoren und dem verfügbaren Hauptspeicher) für diesen Job allokiert. Der Benutzer muss explizit festlegen, wie viele Knoten er will und welcher Knotentyp für seinen Job erforderlich ist.

Ab der Version 7.0 und höher unterstützt PBS Professional die enge Integration verschiedener MPI Varianten wie MPICH oder LAM mit dem Lastverteilungssystem. Die parallelen Jobs können über eine PBS-Variante des mpi_run Kommandos gestartet werden, welches dafür sorgt, dass eine Ankopplung

der MPI Tasks auf den einzelnen Knoten an den Taskmanager des PBS-MOM demons stattfindet. Diese Ankopplung stellt sicher, dass PBS die MPI Jobs starten und verwalten kann, ohne dass eine manuelle Intervention des Anwenders oder Administrators erforderlich ist. Insbesondere ist eine automatische Reaktion für alle Fehlerfälle (z.B. „Aufräumen“ der Knoten bei Jobabsturz, etc.) ohne Zuhilfenahme externer Skripte gewährleistet.

Mittels dieser neuen Funktion ist es auch möglich, ein exaktes Accounting auf allen involvierten Rechnerressourcen durchzuführen.

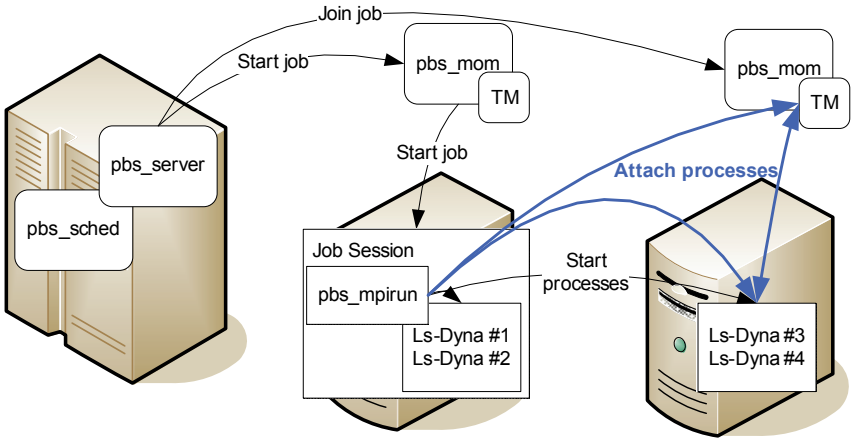

Bild 7: Management von MPI Jobs

5.3 Sonstige Features

Ein weiterer wichtiger Qualitätsaspekt eines Lastverteilungssystems ist die Frage, inwieweit spezielle Features einzelner Hardware Hersteller unterstützt werden. Altair Engineering unterhält strategische Partnerschaften zu allen wichtigen Hardwareherstellern wie Cray, Dell, hp, IBM, NEC oder auch sgi. Zu den von PBS Professional 7.0 unterstützten Features zählen u.a.:

- *CPU-Sets auf SGI Altix Systemen*
- *Comprehensive System Accounting auf SGI Altix*
- *Job Containers auf SGI Altix*
- *Checkpoint/Restart auf Systemlevel unter IRIX und UNICOS*
- *Integration mit IBM MPI on AIX 5 auf POWER4 & 5 mit IBM’s Parallel Operating Environment (POE)*
- *Integration mit CRAY’s “psched” Utility*
- *Integration mit SUN’s HPC Toolkit.*

Eine Zusammenfassung der von PBS unterstützten Plattformen gibt die nachstehende Abbildung:

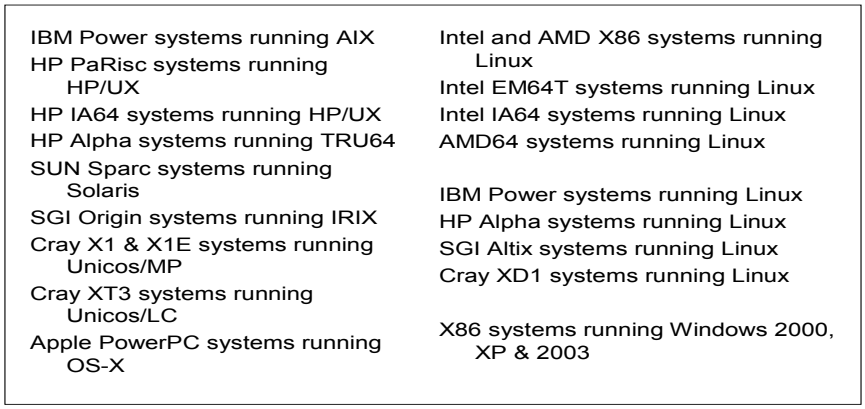

Bild 8: Von PBS unterstützte Plattformen

Zusammenfassung und Ausblick

PBS Professional hat sich als intelligentes Lastverteilungssystem mit umfassender Funktionalität am Markt etabliert und wird auf einer Vielzahl von Industriesegmenten und Hardwareplattformen eingesetzt, bis hinauf zu den größten Systemen in der TOP 500 Liste. Altair Engineering ist strategischer Partner aller wichtigen Hardwareanbieter, beispielsweise ist PBS Professional eine wichtige Komponente in HP's Unified Cluster Portfolio ebenso wie in der „Grid & Grow" Initiative von IBM.

Vor kurzem haben Altair Engineering und Scali eine enge Kooperation angekündigt, mit dem Ziel, PBS Professional in die Scali Manage Lösung zu integrieren. Die integrierte Lösung ermöglicht es den Kunden, ihre HPC-Infrastruktur und ihre Server Workloads von einer einheitlichen Management Plattform aus kostengünstig und effizient zu verwalten. Die Vereinbarung schließt auch eine wechselseitige Vermarktung der jeweiligen Softwareprodukte mit ein; Scali wird PBS Professional im Rahmen der Scali Manage Software mit anbieten und dafür auch technische Unterstützung leisten. Andererseits haben PBS Professional Kunden damit die Möglichkeit, eine komplette Cluster Management Lösung aus einer Hand zu bekommen.

Im Rahmen der zunehmenden Nutzung von Gridanwendungen im kommerziellen Umfeld stellt PBS durch seine offene Architektur eine attraktive Ausgangsplattform dar. Altair Engineering stellt seinen Kunden auf Wunsch den Sourcecode von PBS zur Verfügung, so dass spezifische Entwicklungen und Anpassungen an offene Gridumgebungen möglich werden.

Durch das einfache und kostengünstige Lizensierungsmodell, die umfassende Funktionalität und Robustheit der Software sowie die ausgewiesene Lösungskompetenz von Altair im CAE-Umfeld stellt PBS Professional für CAE-Anwender eine hervorragende Lösung für Lastverteilungsaufgaben dar.

Ansprechpartner:

- *Dr. Jochen Krebs, Business Development,* krebs@altair.de *, Tel. +49-171- 3357722*
- *Dr. Ralf Eichmann, Technischer Support,* eichmann@altair.de*, Tel. +49-7031-6208-39*

Mehr Informationen unter:

- *http://www.altair.com*

