Parallele Hochleistungs-Ein-/Ausgabe in Cluster-Umgebungen

Parallele und verteilte Systeme
Institut für Informatik
Universität Heidelberg
Prof. Thomas Ludwig

pvs.informatik.uni-heidelberg.de

Forschungsgebiete an der TUM

- Lastausgleich durch Prozeßmigration
  - Auf Parallelrechnern (Intel iPSC)
  - Auf Rechner-Clustern
- Werkzeuge-Infrastrukturen
  - Monitoringsysteme für parallele Debugger, parallele Leistungsanalyse (selber wieder paralleles Programm)
  - Schnittstellenspezifikation für interoperable parallele Werkzeuge

Forschungsgebiete in Heidelberg

Parallele Hochleistungs-Ein-/Ausgabe für Cluster-Umgebungen

Warum ist E/A kritisch?

- Heidelberger Helics-Cluster (Rang 165 weltweit) mit 256 Knoten à 2GByte Hauptspeicher erzeugt z.B. 0.5TByte Daten pro Iterationsschritt einer numerischen Simulation
- Ergebnisdaten heute im Bereich TByte-PByte
- Gebiete: Bioinformatik, Physik u.a.

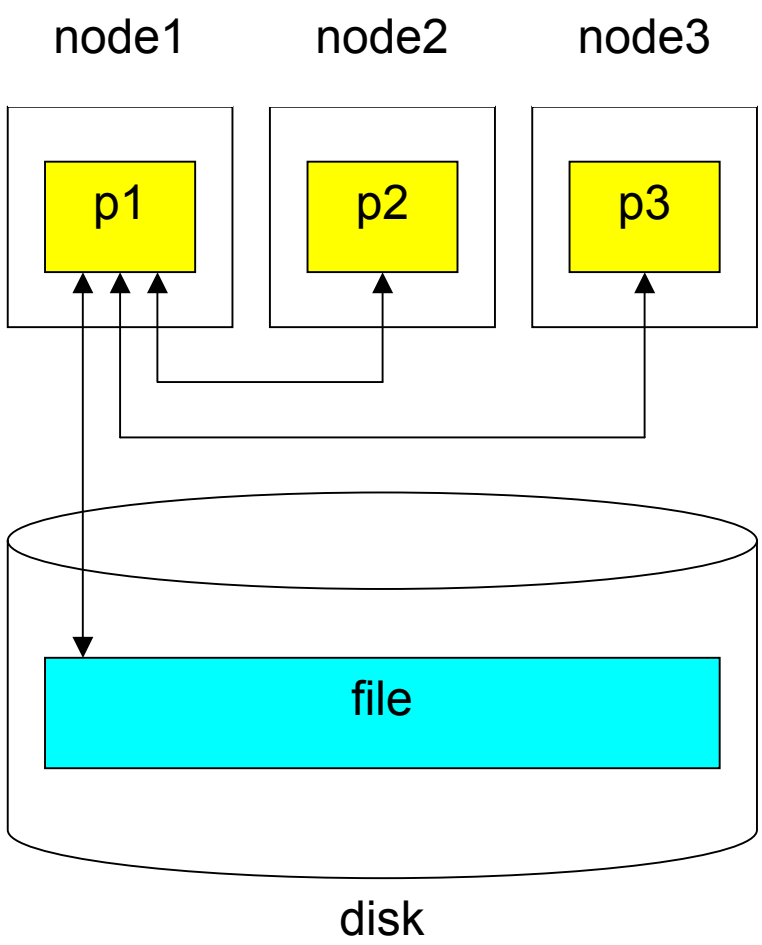
Parallele Ein-/Ausgabe

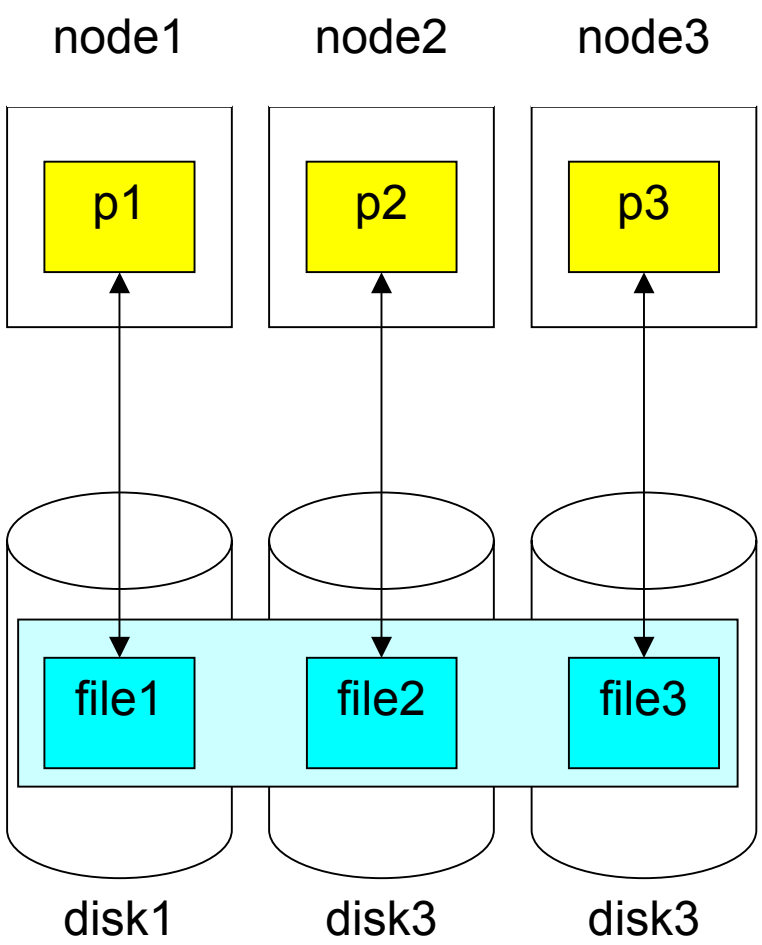
Parallele Ein-/Ausgabe auf zwei Ebenen

- Auf der Programmebene
  - Viele parallele Prozesse schreiben in dieselbe Datei
  - Zugriff zu verschiedenen/identischen Positionen
  - Disjunkte/nichtdisjunkte Sicht auf die Dateien
- Auf der Systemebene
  - Datei ist physikalisch über viele Platten verteilt
- Typisches Abbildungsproblem; Lokalität wichtig

Man muß nicht beide Ebene haben!

Nichtparallele vs. parallele E/A





Parallele Ein-/Ausgabe...

Ansätze

- Auf der Programmebene
  - MPI-IO (Bestandteil von MPI2)
    Definiert API für parallele E/A
    Syntax + Zugriffssemantiken
- Auf der Systemeben
  - GPFS (IBM)
  - Lustre (Cluster File Systems)
  - PVFS (Argonne National Labs / Clemson University)

Fragestellungen auf Benutzerebene

- Welche Zugriffssemantik eignet sich für welche Anwendung?
  (gemeinsamer/individueller Dateizeiger, gemeinsame/individuelle Datenbereiche)
- Wie kann ich meine E/A charakterisieren?
  (zusammenhängend, zufällig, kleine/große Datenmengen)
- Wie kann ich die E/A-Leistung messen?
- Wie kann ich meinen Datendurchsatz optimieren?

Fragestellungen auf Systemebene

- Wie werden die Daten einer Datei über die Platten hinweg verteilt?
- Wie mißt man hier die Lasten im System?
- Automatische Lastoptimierung durch Verlagerungen von Dateiteilen?
- Wie optimiert man Zugriffe auf die Metadatenserver? (Zentrale Komponente als Engpaß im System)
- Was passiert beim Absturz des Programms und beim Ausfall von Platten?
- Wie übertrage ich Riesendateien auf andere Systeme? (Datensicherung)
- Wie verwende ich Riesendateien in anderen Programmen? (Datenvisualisierung)

Laufende Arbeiten

Steuerbare Ressourcennutzung in parallelen Dateisystemen

- Konzept
  - Abbildungsfunktion von logischer Dateiposition auf physische Dateiposition wird dynamisch gesteuert
- Ergebnis
  - Lastausgleich durch Verlagerung von Dateiteilen im parallelen Dateisystem
- Nutzen
  - Bessere Ressourcenauslastung
  - Flexiblere Ressourcennutzung

Laufende Arbeiten...

Optimierung der Leistung der Metadatenserver in parallelen Dateisystemen

- Problem
  - Zugriffe auf Dateien erfordern vorherige Zugriffe auf Metadaten dieser Datei
  - Konzepte mit zentralem Metadatenserver skalieren nicht
- Ansatz
  - Replikation der Metadatenserver
  - Einsatz von Caching- und Locking-Mechanismen
- Wichig: Benchmark für Metadatenserver

Laufende Arbeiten...

Parallele E/A in numerischen Programmen

- Umstellung relevanter Programme auf MPI-IO
- Evaluierung unserer neuen Verfahren

Forschungslandschaft parallele E/A

- In Deutschland nahezu keine Forschung hierzu
  - Karlsruhe: Tichy / Isaila
    Clusterfile (noch nicht frei verfügbar)
- Weltweit auch wenige Gruppen