Netzwerkunterstützung für Grid Computing

("Grid networking, ned Gridworking!")

Michael Welzl http://www.welzl.at

DPS NSG Team http://dps.uibk.ac.at/nsg
Institut für Informatik
Universität Innsbruck

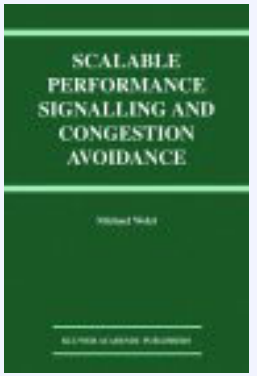
OVE / OCG Meeting
Wien
11. October, 2005

Überblick

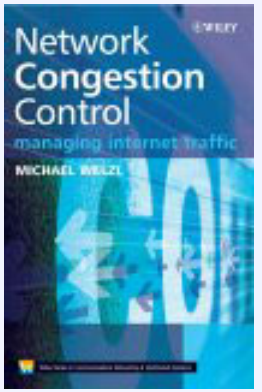
- Vorstellung; das NSG Team an der Universität Innsbruck
- Thematische Eingrenzung und Problemstellung
- Verbesserungsansätze
  - Beispiel 1: Netzwerkmessung
  - Beispiel 2: Dienstgüte-Garantien und High Performance Kommunikation
- Zusammenfassung

Wer bin ich?

- Ein echter Globetrotter :) Innsbruck ⇒ Linz ⇒ Innsbruck



- Doktorat in Darmstadt (Max Mühlhäuser + Jon Crowcroft)
  - Verteidigung im November 2002 mit Auszeichnung bestanden
  - Veröffentlichung im August 2003 als Kluwer (jetzt Springer) Buch "Scalable Performance Signalling and Congestion Avoidance"
  - Erhielt Preis für beste Dissertation 2004 von GI/ITG KuVS



- Network Congestion Control: Managing Internet Traffic
  - John Wiley & Sons, Juli 2005
  - Das erste einführende Buch zu diesem Thema

- Forschungsannahme: one-size-fits-all TCP + IP nicht optimal
  - Hauptinteresse: Maßschneidern von Netztechnologie für
    - heterogene Infrastruktur (z.B. Hochgeschwindigkeits- oder verrauschte Verbindungen, mobile Einsatzszenarien)
    - heterogene Anwendungen (z.B. Streaming media, Signalisierung, Grid)

Das NSG Team

Werner Heiss
Tiroler Wissenschaftsfonds

Murtaza Yousaf
Stipendium der Pakistanischen Regierung

Michael Welzl
Institut für Informatik

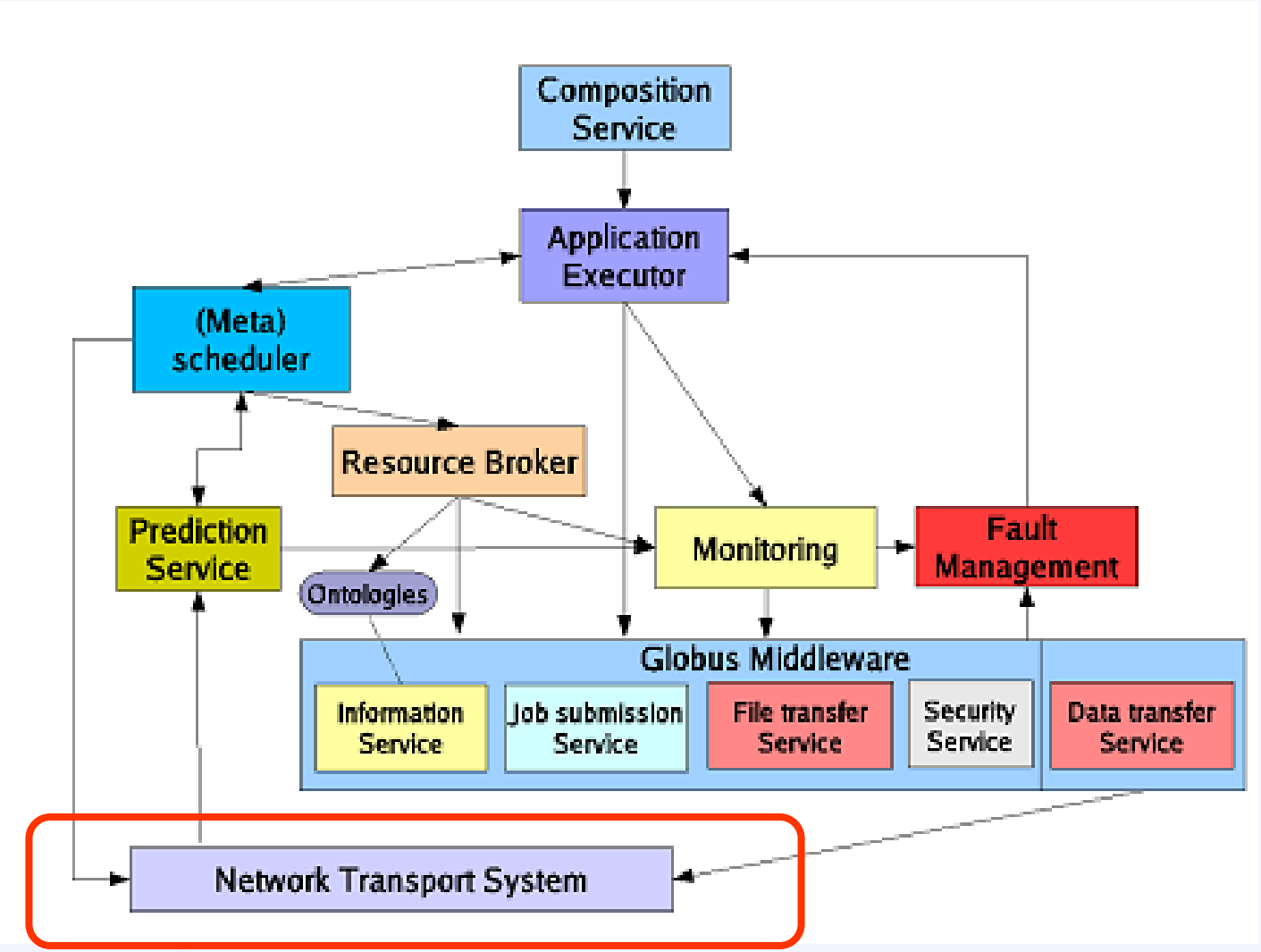
Sven Hessler
FWF

Nicht im Bild- beginnt November 2005: **Kashif Munir**
Stipendium der Pakistanischen Regierung

NSG Tätigkeiten

- Forschungsthemen: Grid = Hauptfokus
  - Maßgeschneiderte Netzwerktechnologie für Grid Anwendungen
    - Staukontrolle
    - Quality of Service (QoS)
    - Transportprotokolle
    - Netzwerkmessungen und Vorhersagen
    - Middleware Kommunikation
  - Weiters: andere Aspekte von Rechnernetzen (z.B. Multimediakommunikation)
- Lehre: wir decken die Rechnernetz-Lehre an der Universität Innsbruck ab
- Kooperationen: Grid-spezifische Ergebnisse werden...
  - mittels der GHPN-RG des Global Grid Forum (GGF) zu Standards beigetragen
  - in das Workflow System eingebettet, das von der DPS Arbeitsgruppe an der Universität Innsbruck entwickelt wird

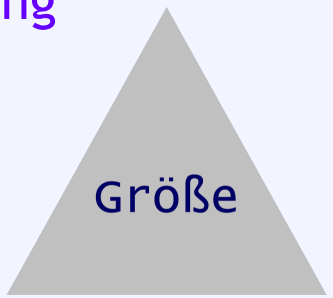
Das DPS Grid Workflow Application Execution Environment



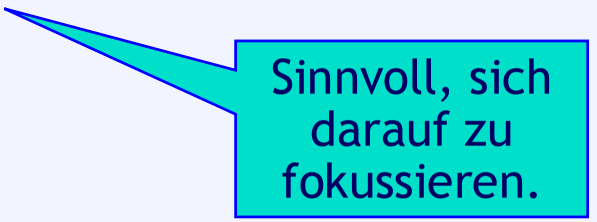
Thematische Eingrenzung und Problemstellung

Thematische Eingrenzung

- Grid Geschichte: paralleles Rechnen in wachsendem Umfang
  - Parallele CPU Architekturen
  - Multiprozessor Maschinen
  - Cluster
  - ("Massiv verteilte") Rechner am Internet



- Herkömmliches Ziel: Rechenpower
  - Gridforscher = „Parallele Systeme" Forscher; daher hat sich das Ziel nicht stark verändert



- Breitere Definition ("Ressourcenzugriff")
  - nötig - z.B. haben Computer auch Festplatten :-)
  - Neue Forschungsgebiete / Schlagwörter : Wireless Grid, DataGrid, Pervasive Grid, [*dieser Platz ist reserviert für Ihr Forschungsgebiet*] Grid
  - manchmal vielleicht etwas *zu* breit, z.B. gehört die "P2P Working Group" jetzt zum Global Grid Forum

Grid Anforderungen

- Effizienz + leichte Bedienbarkeit !
  - Programmierer sollte sich nicht um das Grid „kümmern“ müssen
  - Anwendungen sollten automatisch verteilt werden
- Darunter liegendes System muss sich kümmern um
  - Fehlerbehandlung
  - Authentifizierung, Autorisierung and Verrechnung
  - Effizientes Scheduling / Lastverteilung
  - Finden und Zuteilen von Ressourcen
  - Benennung
  - Ressourcenzugriff und Monitoring
- Kein Problem: das machen wir alles - in Middleware
- de facto Standard: “Globus Toolkit“
  - Installation von GT3 in unserem „High Performance System“: ca. 1 1/2 Stunden
  - ja, es macht wirklich alles :) 1000e Erweiterungen - MDS, NWS, GRAM, ..

Problem: wie Gridforscher das Internet sehen

- Abstraktion - einfach verwenden was verfügbar ist
  - dennoch: Performance = Hauptziel

Genau wie die Web Service Community

Konflikt!

- Existierendes Transportsystem
  (TCP/IP + Routing + ..) funktioniert gut

Absolut nicht wie die Web Service community !

- QoS macht alles besser, das Grid braucht es!
  - und dank IPv6 gibt's da eine Chance

Falsch.

- Zitat aus einem Kommentar eines Gutachters für einen Artikel:
  "In fact, any solution that requires changing the TCP/IP protocol stack is practically unapplicable to real-world scenarios, (..)."

- Wie man das verändert: GGF GHPN-RG
  - Dokumente wie "net issues with grids", "overview of transport protocols"
  - außerdem einige EU Projekte, Workshops, ..

Grid-Netz Eigenheiten

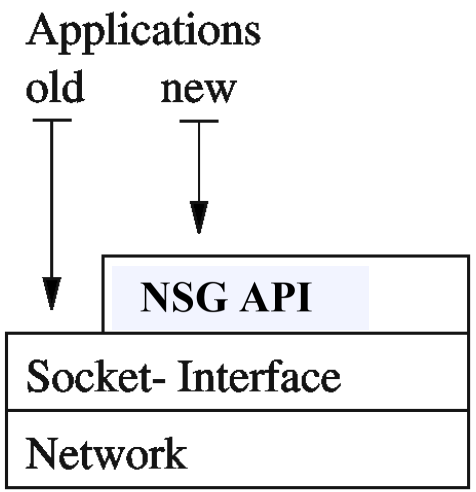
- Besonderes Verhalten
  - Vorhersagbares Verkehrsaufkommen - das ist vollkommen neu für das Internet!
    - Web: Benutzer erzeugen Verkehr
    - FTP download: beginnt... endet
    - Streaming video: entweder CBR oder abhängig von Inhalt! (Kopfbewegung, ..)
      - Könnte von Staukontrollverfahren genützt werden
  - Unterscheidung: „Bulk" Dateitransfer (z.B. GridFTP) vs. Kontrollnachrichten (z.B. SOAP)
  - Dateien werden oft verschickt und nicht angefragt („push" statt „pull")

- Besondere Anforderungen
  - Vorhersagen
  - Zeitschranken, Bandbreitengarantien ("advance reservation") => QoS

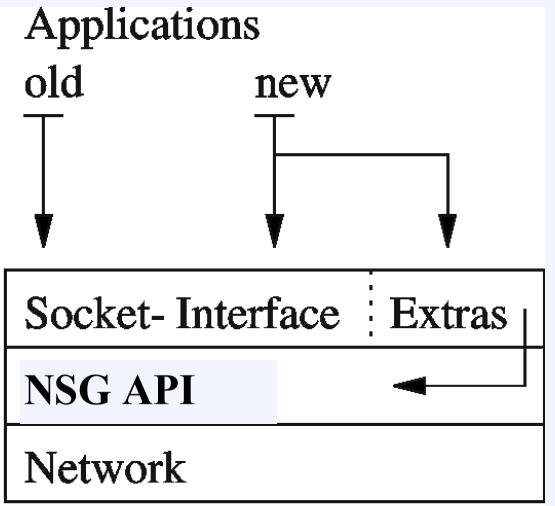
- Verteiltes System, aktiv für eine gewisse Zeit
  - Overlay basierte Strategien können verwendet werden (wird in P2P Systemen gemacht!)
    - Multicast
    - P2P Paradigma: „übernimm Arbeit für andere um das Gesamtsystem zu verbessern" (in Deinem eigenen Interesse) - z.B. transcoding, als PEP agieren, ..
  - Ausgeklügelte Messverfahren können angewandt werden
    - Einige davon brauchen viel Zeit, einige benötigen eine verteilte Infrastruktur

Einige Probleme: Anwendungsinterface...

- Wie spezifiziert man Eigenschaften und Anforderungen?
  - Sollte einfach und flexibel sein - QoS Spezifikationssprachen verwenden?
  - Sollten Anwendungen das überhaupt mitbekommen?
    ⇒ Trade-off zwischen Dienstgüte und Transparenz!

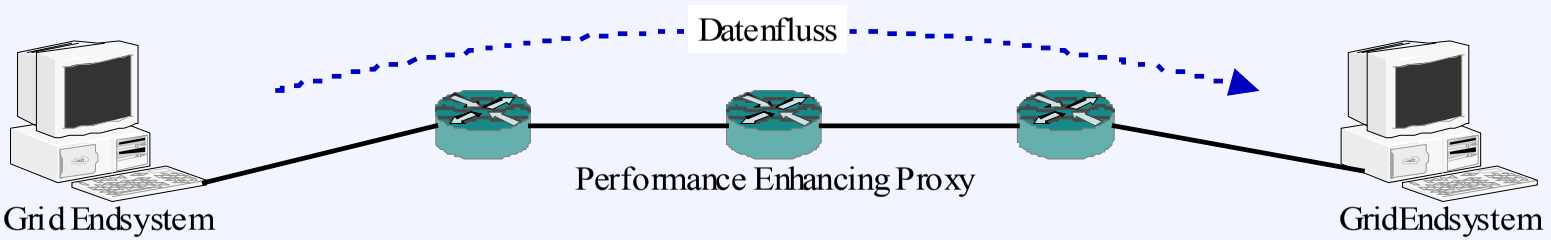


Herkömmliche Methode

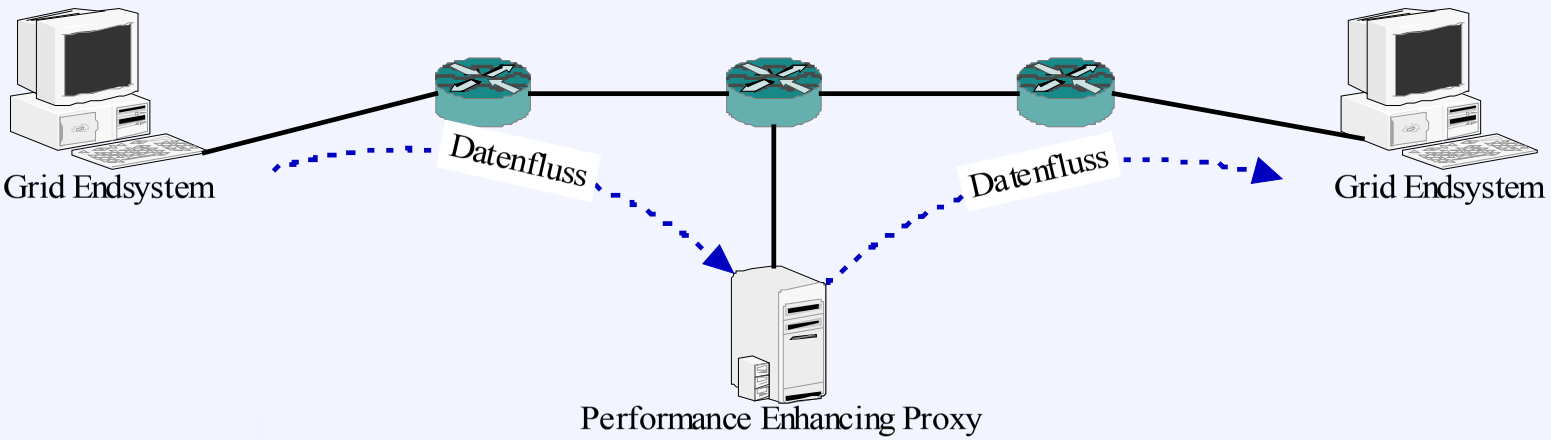


Unser Ansatz

... und Wissen über andere Rechner



(a) Herkömmlicher PEP



(b) NSG PEP

Verbesserungsansätze

Beispiel 1: Netzwerkmessung

Das Netz vermessen

- Wenn man misst, misst man die Vergangenheit
  - Vorhersagen / Schätzungen mit einer ??% Erfolgschance
- Wenn man misst, verändert man das System
  - z.B. UDP vs. TCP: messen ohne Einfluss zu nehmen kann sehr wichtig sein!
- Messungen liefern keine Garantien
  - Internetverkehr = Ergebnis des Benutzerverhaltens!
- Forschung oft in kontrollierbaren, isolierten Umgebungen
- Feldversuche sind ein nötiges Extra wenn man annimmt dass etwas funktioniert

NWS: Der Network Weather Service

- Verteiltes System bestehend aus
  - Name Server (uninteressant)
  - Sensor - eigentliche Messinstanz, speichert regelmäßig Werte in.....
  - Persistent State
  - Forecaster (Berechnungen auf den Daten im Persistent State)

- Interessante Elemente:

  - Sensor
    Gemessene Ressource: availableCpu, bandwidthTcp, connectTimeTcp, currentCpu, freeDisk, freeMemory, latencyTcp

  - Forecaster
    Wendet unterschiedliche Modelle zur Vorhersage an, vergleicht mit den eigentlichen Messdaten, verwendet in Zukunft das Modell, das am besten funktioniert hat

Dauer einer langen TCP Übertragung

RTT einer kleinen Nachricht

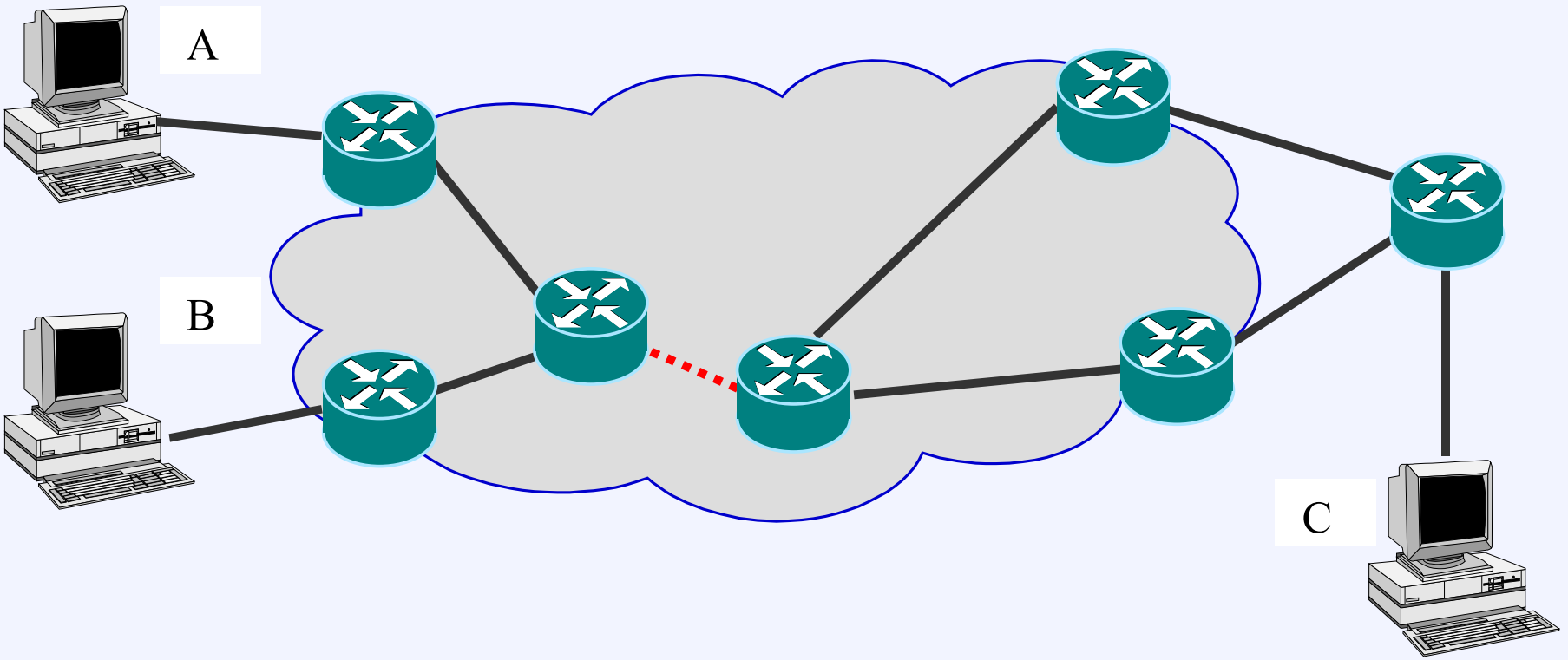
Kritik an NWS

- Architektur (Aufteilung in Sensoren, Forecaster usw.) scheint vernünftig; open source ⇒ vielleicht sinnvoll, neue Arbeiten in NWS zu integrieren

- Sensor
  - aktive Messungen obwohl Messen ohne Einflussnahme ein wichtiges Designziel war - NWS verzichtet auf passives messen von TCP (d.h. es ignoriert vorhandene Daten!)
  - Seltsame Messmethode:
    z.B. für „Large message throughput": *"Empirically, we have observed that a message size of 64K bytes (..) yields meaningful results"*
  - ignoriert Paketgröße (= Messgranularität!) und Pfadeigenschaften
  - Triviale Methode - viel ausgeklügeltere Verfahren verfügbar (z.B. packet pair - mehr dazu später!)
  - Punkt-zu-punkt Messungen: nutzt nicht die verteilte Infrastruktur

- Forecaster
  - stützt sich auf diese eigenartigen Messungen, wobei über die Verteilung der Ergebnisse nicht viel bekannt ist (man weiß aber einiges über andere Messmethoden!)
  - verwendet recht triviale Modelle

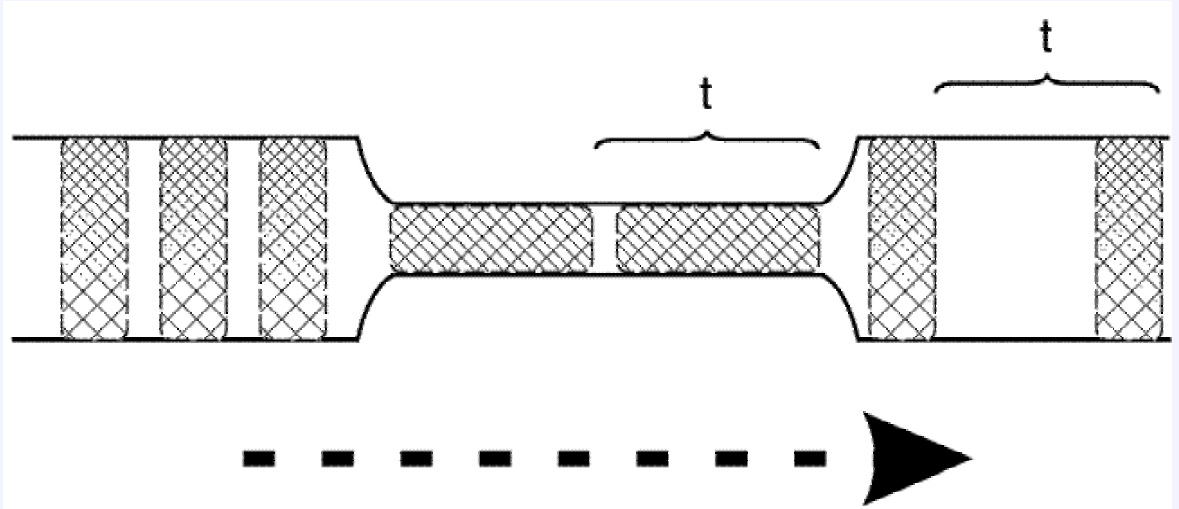
Ausnutzen der verteilten Infrastruktur



- Beispielproblem:
  - C teilt A und B Arbeiten zu (CPU, Speicher verfügbar); beide senden Ergebnisse an C
  - B behindert A - Arbeit von B hätte vielleicht bei C verbleiben sollen!
- Pfadwechsel sind selten - daher möglich, potentielles Problem im Voraus zu erkennen
  - generiere Testnachrichten von A, B zu C - identifiziere Signatur von B in A's Verkehr

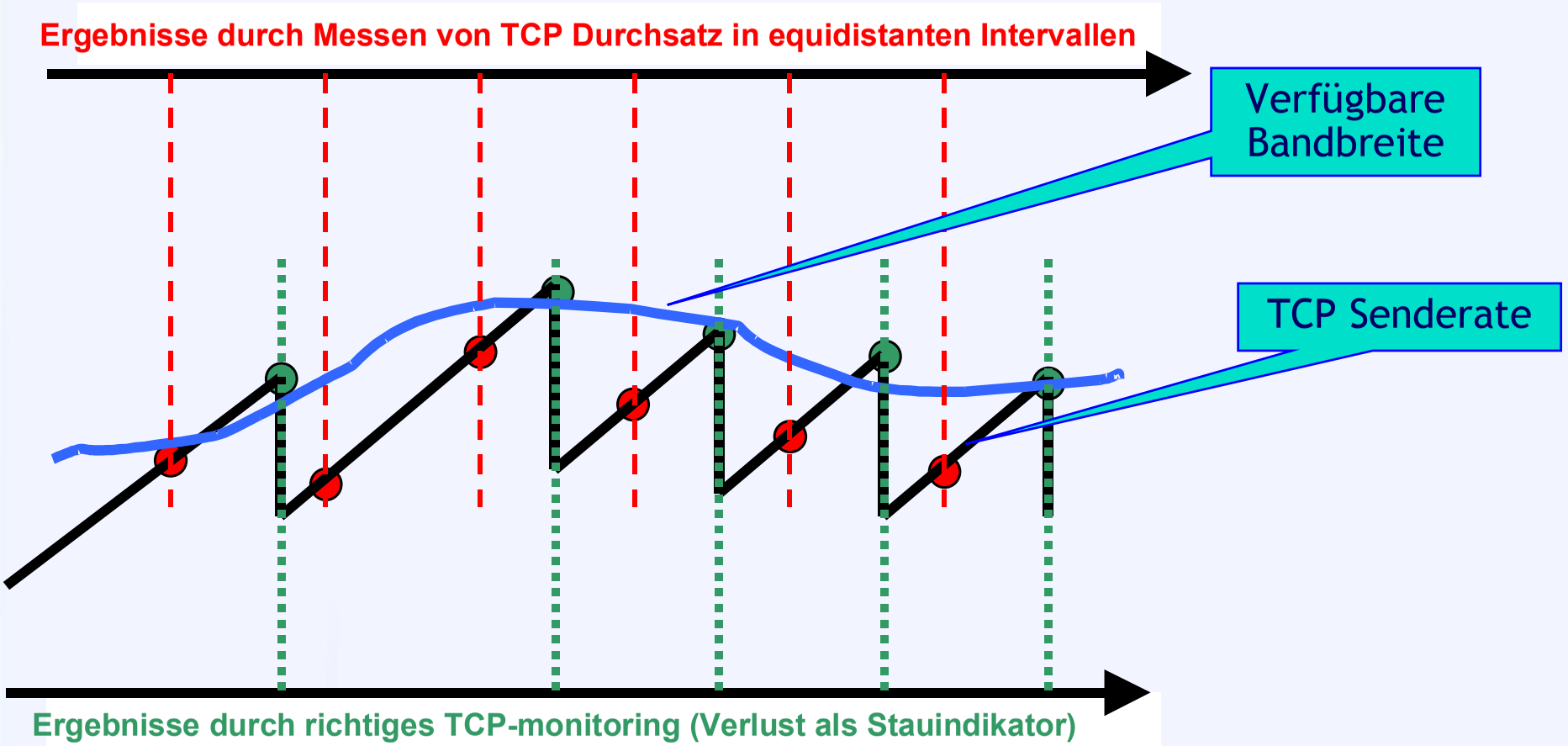
Ausnutzen von Langlebigkeit

- Zeitskala von Verkehrsschwankungen < Zeitskala von Pfadwechseln
  ⇒ Wissen über Verbindungskapazitäten nützlicher als Verkehrsschätzung
- Zugrunde liegende Technik: packet pair
  - sende zwei Pakete p1 und p2 hintereinander; hohe Wahrscheinlichkeit dass p2 genau hinter p1 am Flaschenhals in die Queue gestellt wird
  - Empfänger berechnet Kapazität am Flaschenhals mit Zeit zwischen p1 und p2
  - Minimiere Fehler über mehrfache Messungen
  - TCP mit "Delayed ACK" Empfänger schickt automatisch Paketpaare
    ⇒ passives Beobachten von TCP Empfänger ist nützlich!



Verkehrsvorhersage durch Beobachten von TCP

- TCP überträgt Selbstähnlichkeit am Flaschenhals an Endsysteme
- Automatische Vorhersage? Schwierig, aber möglich, denke ich - z.B.:
  Yantai Shu, Zhigang Jin, Jidong Wang, Oliver W. W. Yang: Prediction-Based Admission Control Using FARIMA Models. ICC (3) 2000: 1325-1329



Beispiel 2: Dienstgüte-Garantien und High Performance Kommunikation

QoS (reservieren von Netzverbindungen),
High Performance Kommunikation für das Grid

QoS: der state-of-the-art :-(

Artikel beim SIGCOMM'03 RIPQOS Workshop: "Why do we care, what have we learned?"

- QoS`s Downfall: At the bottom, or not at all! Jon Crowcroft, Steven Hand, Richard Mortier,Timothy Roscoe, Andrew Warfield
- Failure to Thrive: QoS and the Culture of Operational Networking Gregory Bell
- Beyond Technology: The Missing Pieces for QoS Success Carlos Macian, Lars Burgstahler, Wolfgang Payer, Sascha Junghans, Christian Hauser, Juergen Jaehnert
- Deployment Experience with Differentiated Services Bruce Davie
- Quality of Service and Denial of Service Stanislav Shalunov, Benjamin Teitelbaum
- Networked games --- a QoS-sensitive application for QoS-insensitive users? Tristan Henderson, Saleem Bhatti
- What QoS Research Hasn`t Understood About Risk Ben Teitelbaum, Stanislav Shalunov
- Internet Service Differentiation using Transport Options:the case for policy-aware congestion control Panos Gevros

Hauptgründe für das Versagen von QoS

- Benötigte Teilnahme von Endanwendern und allen dazwischen liegenden ISPs
  - "normaler" Internet Benutzer möchte Internet-weite QoS, oder gar keine QoS
  - In einem Grid, möchte ein "virtuelles Team" QoS zwischen seinen Rechnern
  - Mitglieder des Teams teilen die gleichen ISPs - Geldfluss ist möglich
- Technische Unfähigkeit individuelle (per-flow) QoS anzubieten
  - "normale" Internet Benutzer
    - unbegrenzte Anzahl Datenflüsse kommt und geht zu jeder Zeit
    - heterogener Verkehrsmix
  - Grid Benutzer
    - Anzahl der Mitglieder eines "virtuellen Teams" evtl. begrenzt
    - klare Unterscheidung zwischen „bulk" Dateiübertragungen und Kontrollnachrichten
    - Auftreten von Datenflüssen wird von Maschinen, nicht von Menschen gesteuert
- $\Rightarrow$ QoS könnte für das Grid funktionieren !

High Performance Kommunikation

- Oft werden große Dateien in Grids übertragen und dafür Verbindungen mit hoher Kapazität gekauft. Daher, zwei Ziele:
  - effiziente Nutzung der Kapazität: wünschenswert 1 gbit/s über eine 1 gbit/s Verbindung zu erzielen
  - Fairness: wenn 10 Datenflüsse eine Verbindung Teilen, sollten alle 10 Datenflüsse ihren Anteil erhalten
    = Effizienz: z.B. sollte GridFTP nicht SOAP Nachrichten blockieren

- Standard seit 1980ern: Transmission Control Protocol (TCP)
  - grob: additives Ratenerhöhen bis Queue am Flaschenhals überläuft und Paketverlust auftritt (Stau erzeugt!), dann Halbieren der Rate ⇒ Sägezahn
  - funktioniert schlecht in den Umgebungen von heute: Hochgeschwindigkeitsverbindungen, "long fat pipes", verrauschte (Funk) Verbindungen, ..
  - schrittweise (kleine + abwärts kompatible) Verbesserungen standardisiert

- Viele Alternativen vorgeschlagen, gelegentlich im Grid Kontext - aber schwer einzusetzen wegen TCP-friendliness

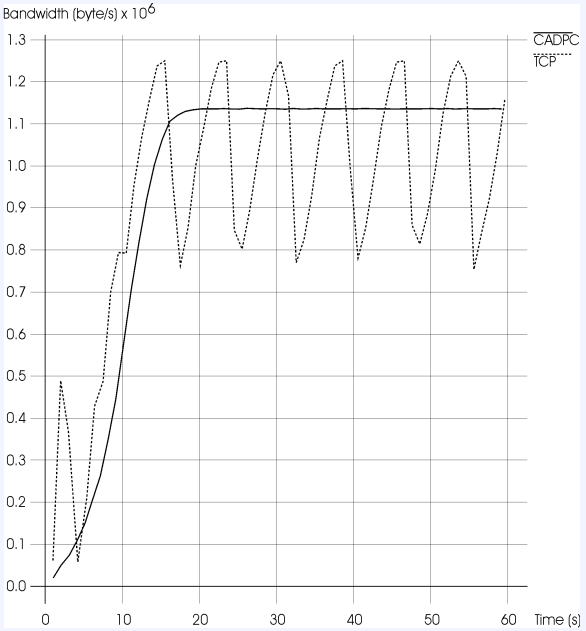
QoS + Staukontrolle = Lösung!

- Idee: verwendete herkömmliches grobgranulares QoS Verfahren (DiffServ) um high-performance „bulk" Dateiübertragung von allem anderen (= SOAP usw. über TCP) zu trennen

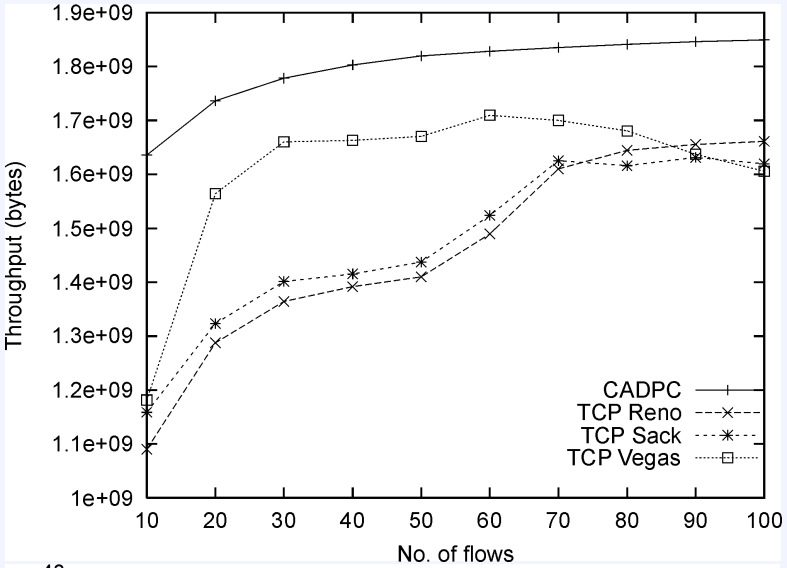
- Isolierter langlebiger Datentransfer = Anforderungen von CADPC/PTP
  - Das ist der beste Staukontrollmechanismus
  - weil ich ihn für meine Dissertation entwickelt habe :-)

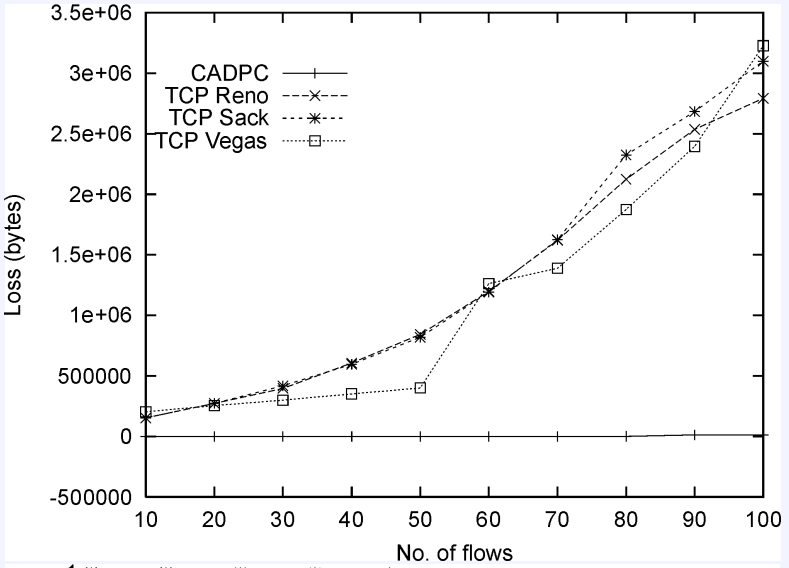
- Einige Eigenschaften:
  - geringer Verlust, hoher Durchsatz
  - vorhersehbare und stabile Rate, die nur von Kapazität und Anzahl der Datenflüsse abhängt

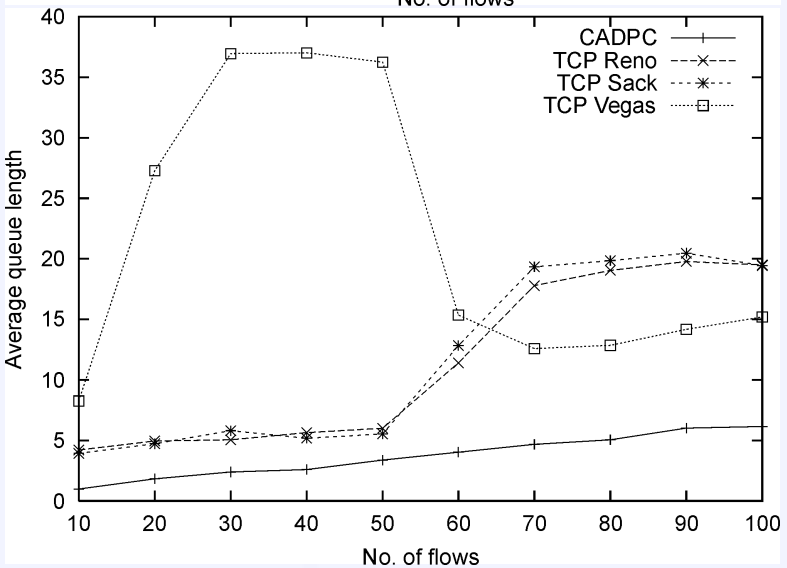
- Nachteil: benötigt Routerunterstützung
  - könnte in einem Grid realistisch sein!

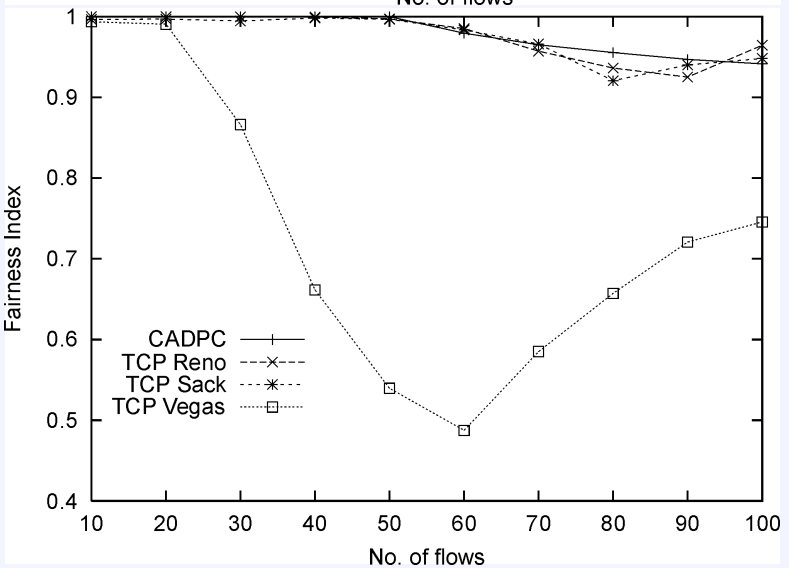


CADPC vs. 3 TCP(+ECN) Varianten

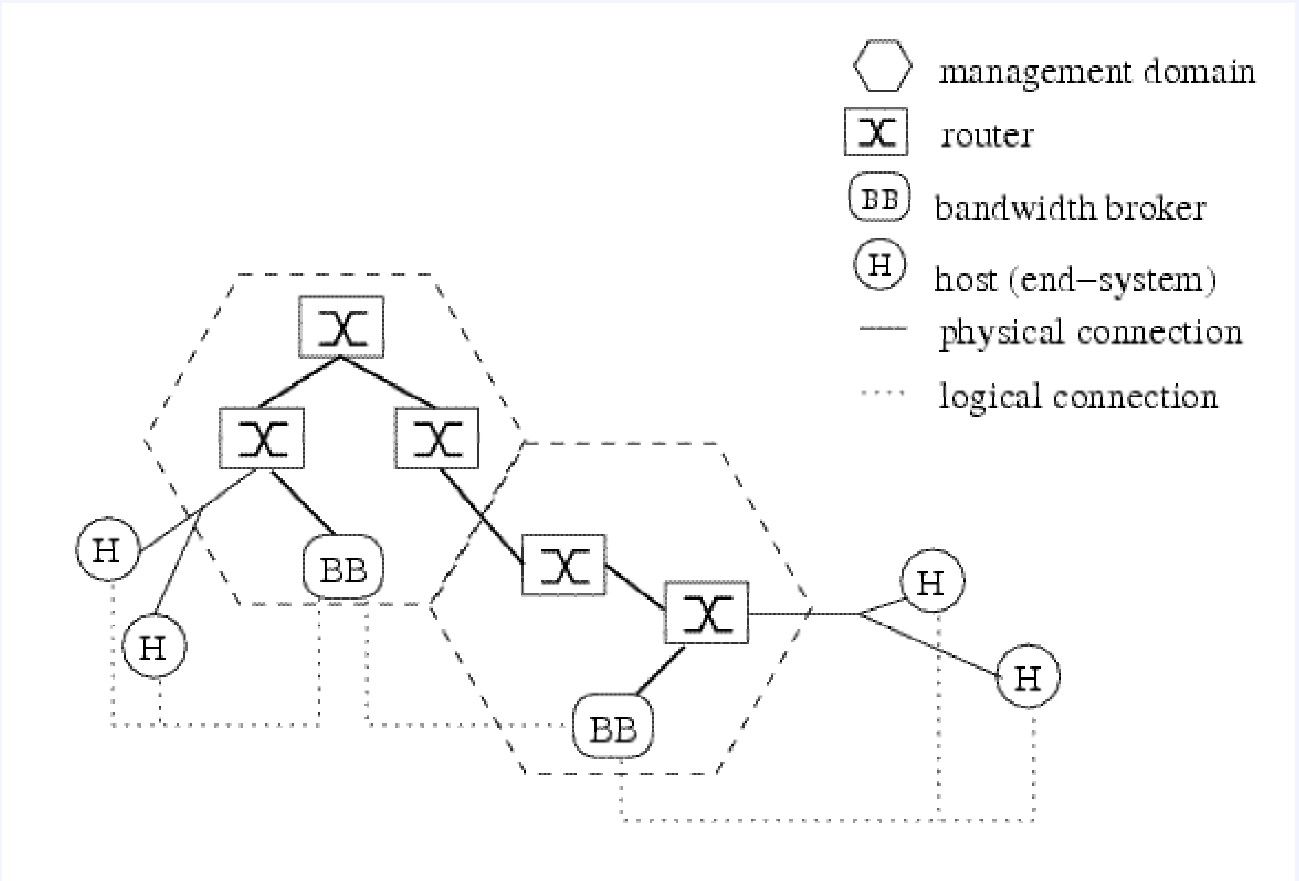








NSG Grid QoS Architektur



- Vorschrift: CADPC/PTP Verwendung für „bulk" Dateiübertragung
- Ressourcenreser-vierung mittels Zugangskontrolle:
  - Bandwidth broker entscheidet was das Netz betreten darf
  - Datenfluss Differenzierung: einfach einem Datenfluss erlauben, sich wie *n* Datenflüsse zu verhalten!

Zusammenfassung

Zusammenfassung

- Grid Anwendungen weisen aus Netzwerksicht besondere Anforderungen und Eigenschaften auf
  - und es ist daher sinnvoll maßgeschneiderte Netzwerktechnologie für sie zu entwickeln.
- Es gibt eine weitere Klasse derartiger Anwendungen...
- Multimedia.
- Für Multimediaanwendungen existiert eine Vielzahl an Netzwerkverbesserungen (einige davon sogar IETF Standards).
- Für das Grid gibt es nichts.
- Das ist eine Forschungslücke; füllen wir sie gemeinsam!

Vielen Dank!

Fragen?