AGC: Ein einigungsbasiertes, rekonfigurierbares Gruppenkommunikationssystem

Hans Reiser
hans.reiser@uni-ulm.de

Abteilung Verteilte Systeme
Universität Ulm

1. Juli 2005

Überblick

Fehlertoleranz in verteilten Systemen

- Verteilte Systeme inzwischen allgegenwärtig
- Gravierende Wirkung von Ausfällen einzelner Teilkomponenten

Frankreich: Handynetz für 20 Stunden lahmgelegt

Millionen Mobilfunkkunden von Bouygues über Netzausfall verärgert 18.11.2004

[...]

Laut Bouygues waren in kürzester Zeit zwei Mammut-Server ausgefallen, von denen der eine im Krisenfall den anderen hätte ersetzen sollen. Firmenchef Gilles Pélisson versprach den Kunden eine Entschädigung.

Computerpanne sorgte für Chaos im Schweizer Weihnachtsgeschäft

Wegen einer Computerpanne ist ausgerechnet im größten Weihnachtsansturm in der ganzen Schweiz der bargeldlose Zahlungsverkehr mit EC-Karten zusammengebrochen. Wütende Kunden stürmten am Samstag ohne ihre Einkäufe aus den Geschäften, berichteten die Schweizer Zeitungen. An den 58.000 Terminals in den Geschäften und fast 5.000 Bargeldautomaten ging nichts mehr. "Eine Katastrophe", kommentierte Bernhard Wenger, Sprecher der Telekurs AG, die den elektronischen Zahlungsverkehr in der Schweiz abwickelt.

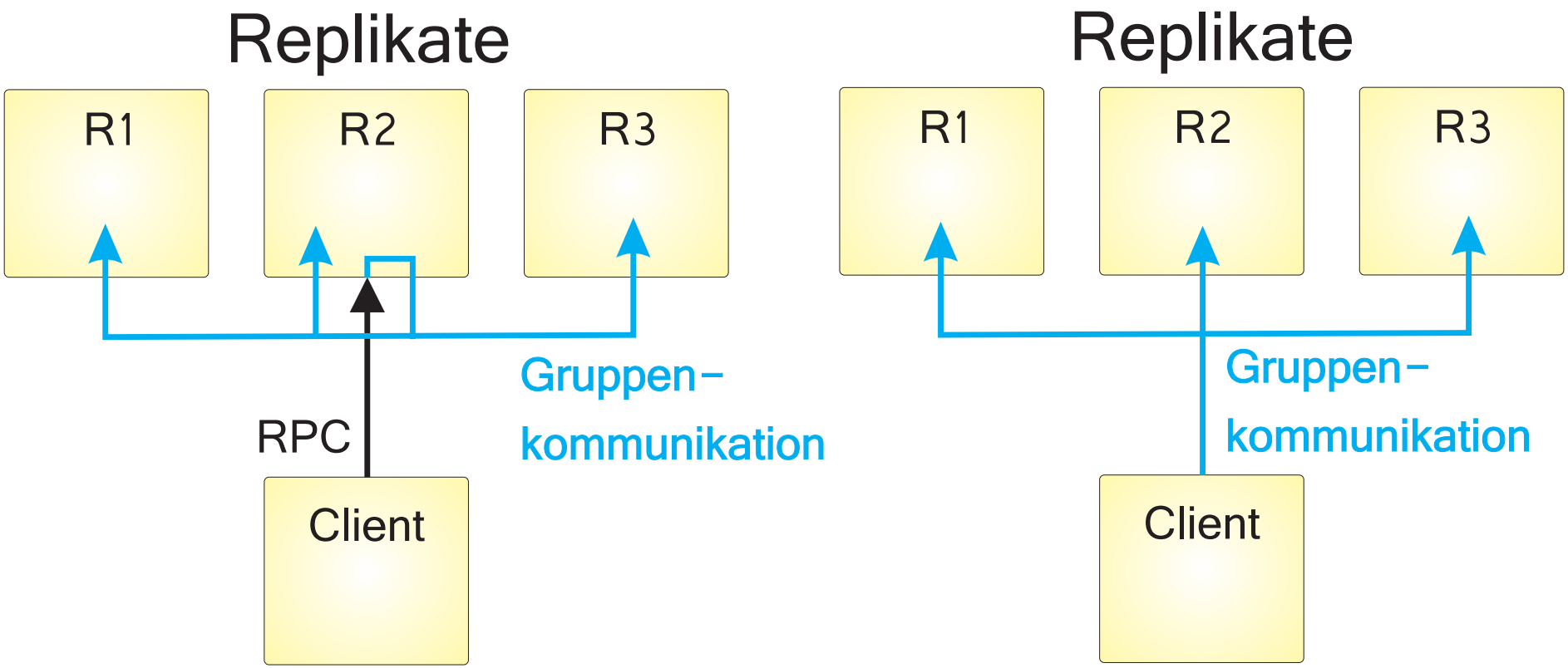
Fehlertoleranz in verteilten Systemen

- Wachsende Abhängigkeit von immer mehr Diensten
  - Dienste für Adressbuch, Terminplaner, etc.
  - Quellcode-Verwaltungssystem
  - Verzeichnisdienste der Infrastruktur (CORBA, RMI, SOAP, ...)
  - usw.

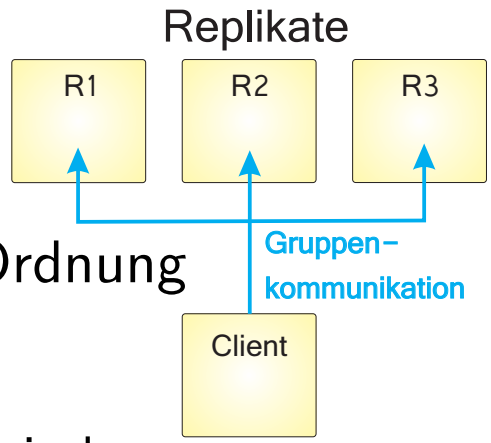
Fehlertoleranz in verteilten Systemen

- Wachsende Abhängigkeit von immer mehr Diensten
  - Dienste für Adressbuch, Terminplaner, etc.
  - Quellcode-Verwaltungssystem
  - Verzeichnisdienste der Infrastruktur (CORBA, RMI, SOAP, ...)
  - usw.
- Redundanz durch Replikation
  - Softwarebasierte Fehlertoleranz (keine teure Spezial-Hardware)
  - Herausforderung: Konsistenz der Replikate
- Universeller Basismechanismus: Gruppenkommunikation
  - aktive Replikation: Zustellung aller Operationen an alle Replikate in einheitlicher Reihenfolge

Aktive Replikation mit Gruppenkommunikation

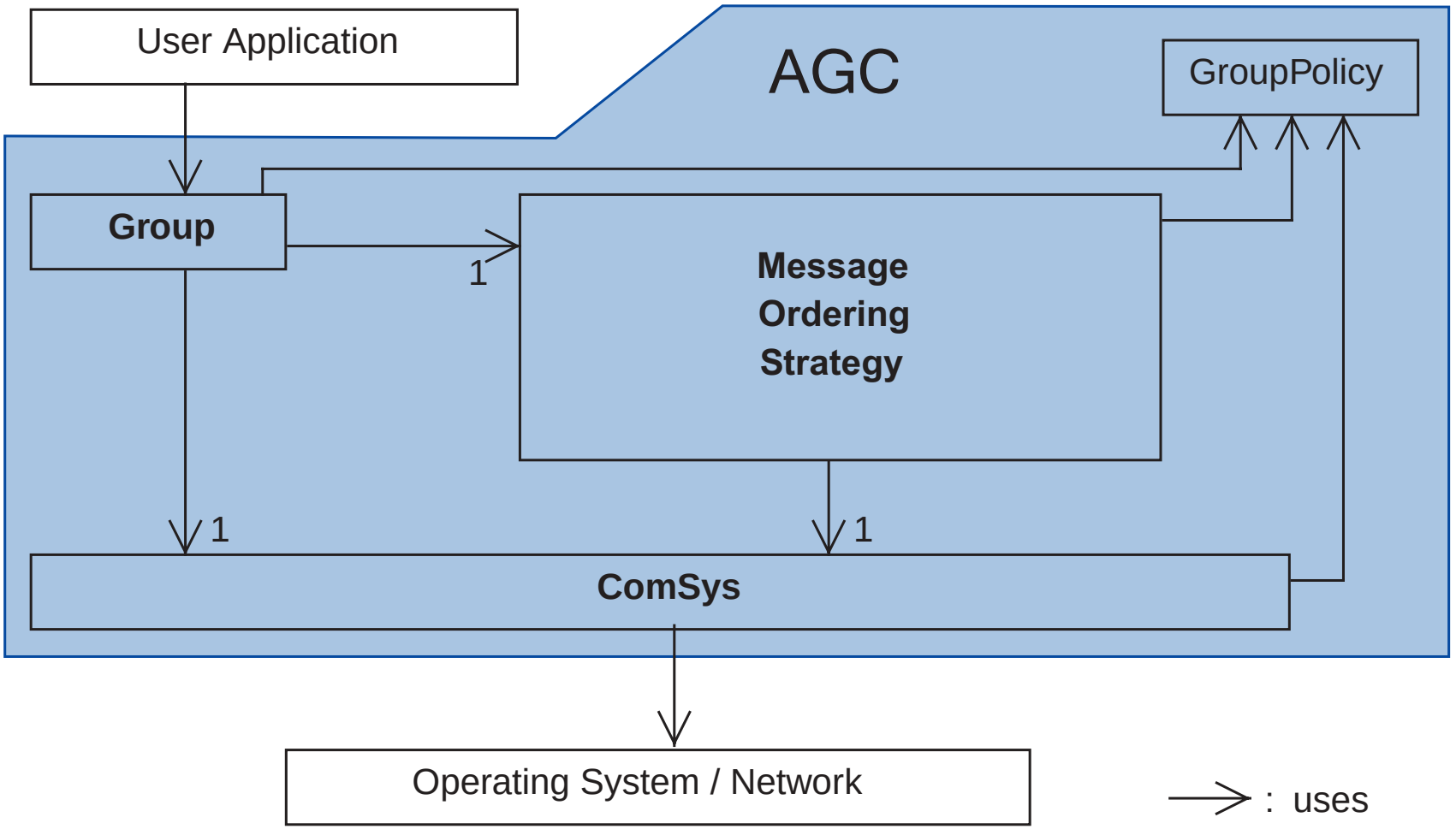


AspectIX Group Communication System (AGC): Ziele

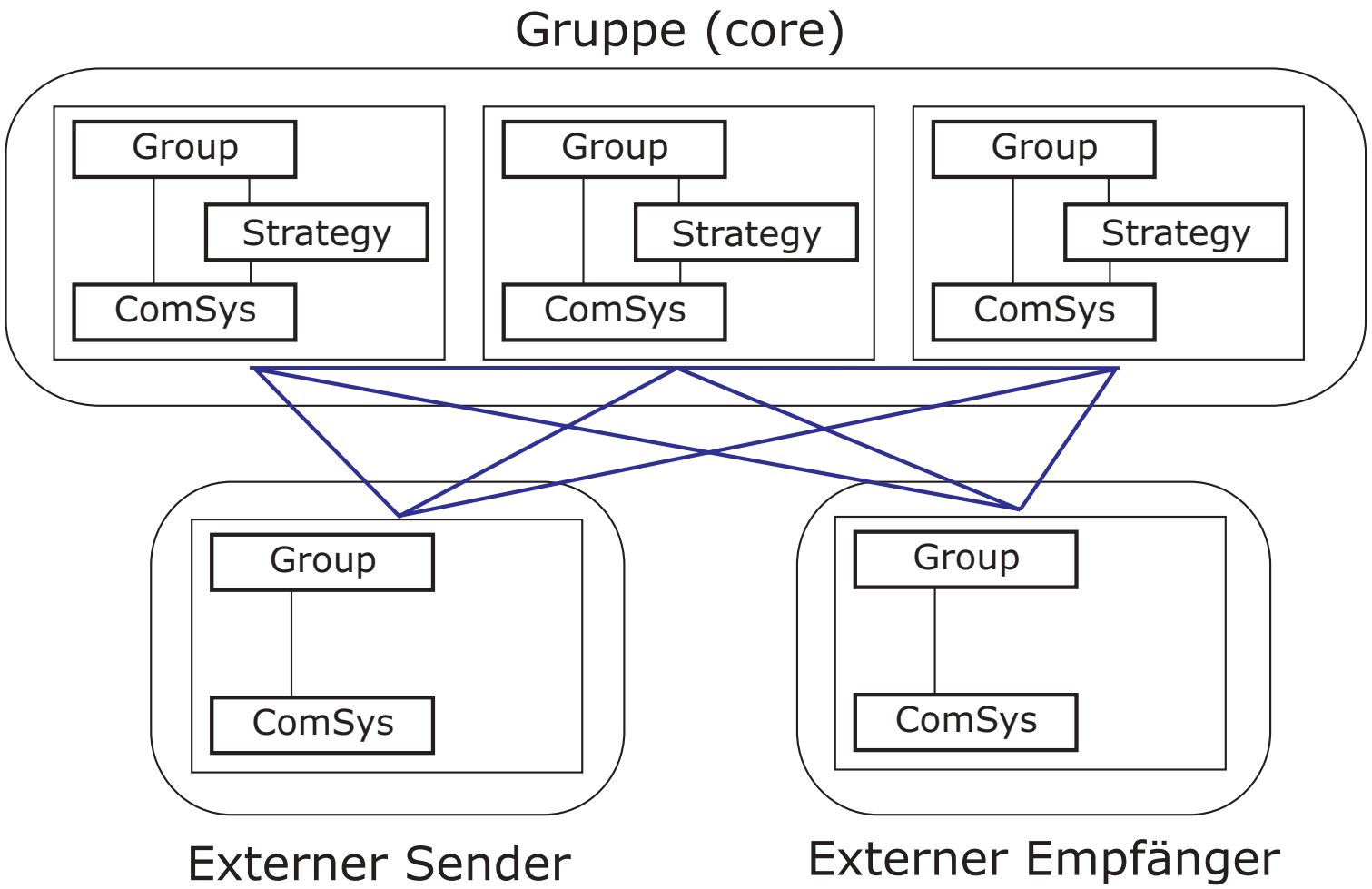


- Zuverlässige Gruppenkommunikation mit totaler Ordnung
- Bedarfsgerechte Mechanismen:
  - Fehlermodelle: crash-stop, crash-recovery, byzantinisch
  - Effizienzoptimierung (Latenz, Durchsatz, Nachrichtenzahl, Recovery)
- Dynamische Rekonfiguration zur Laufzeit
  - Änderung der Gruppenteilnehmer
  - Wechselnde Systemeigenschaften und Anforderungen der Anwendung
  - (Autonome) Anpassung an Nutzungsverhalten der Klienten
- Anwendung: Adaptive Fehlertoleranz in der AspectIX-Middleware

Architektur



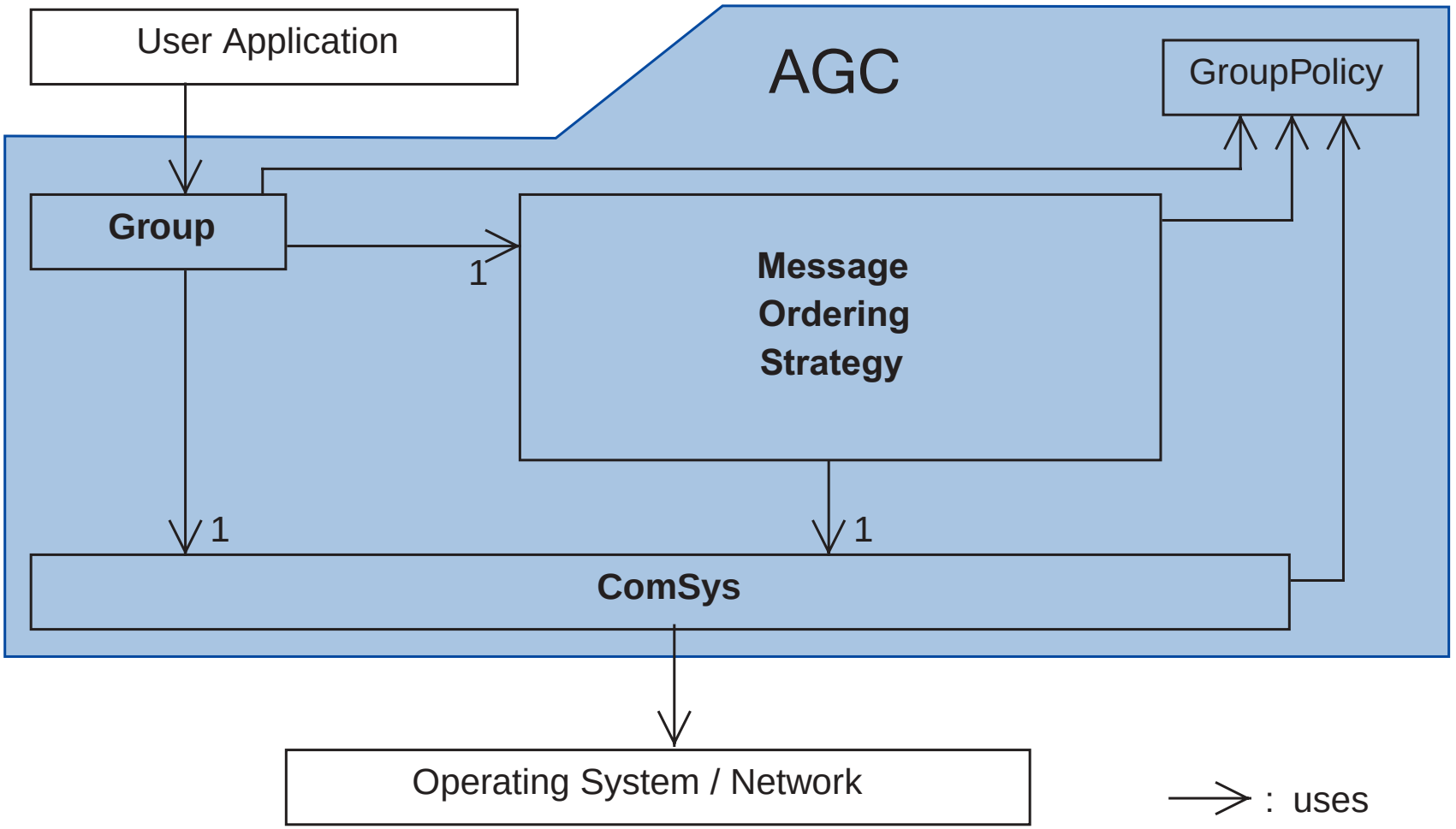
Architektur: Knotentypen



Group-Schnittstelle

Group
join(GroupId group) leave(NodeId node) changePolicy(GroupPolicy p) sendMessage(Message m) sendMsgDirect(Message m, NodeId id) recvMsg(): Message

Architektur



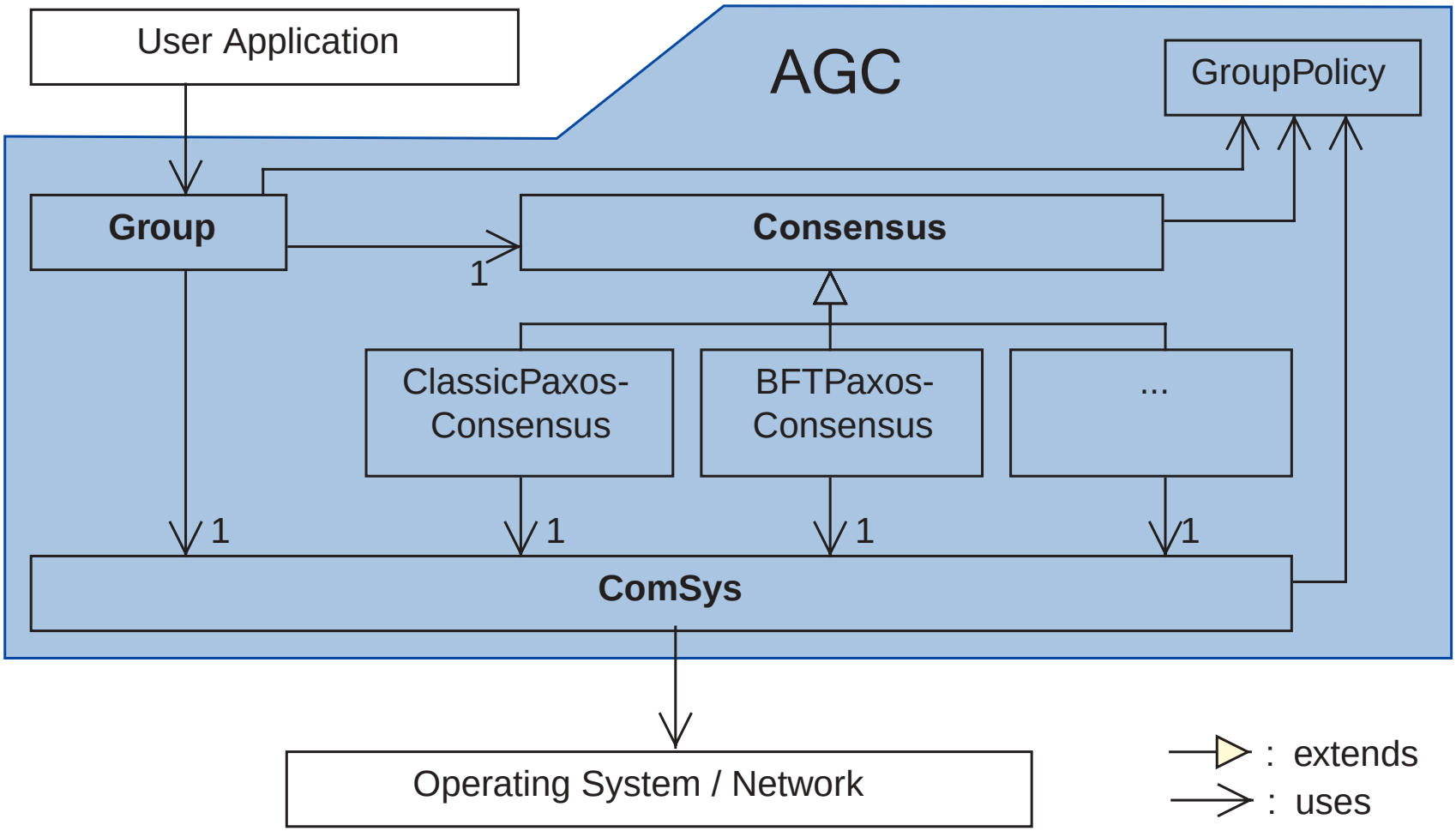
Einigungsbasierte Gruppenkommunikation

- Warum?
  - Effiziente Algorithmen
  - Theoretisch ausgiebig erforscht
  - Nachgewiesene Korrektheit bei geringen Systemanforderungen

Einigungsbasierte Gruppenkommunikation

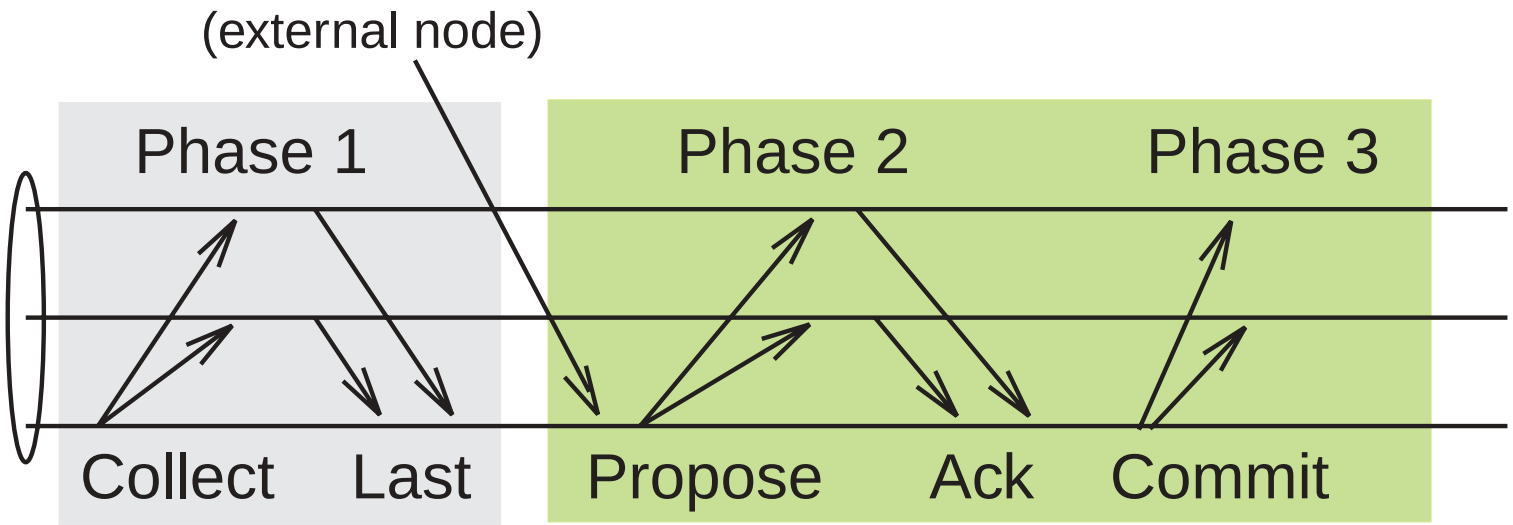
- Warum?
  - Effiziente Algorithmen
  - Theoretisch ausgiebig erforscht
  - Nachgewiesene Korrektheit bei geringen Systemanforderungen
- Implementierte Varianten
  - Klassischer Paxos (crash-stop, crash-recovery)
  - Schneller Paxos (crash-stop, crash-recovery)
  - Byzantinischer Paxos (Castro)

Architektur



Die Paxos-Idee

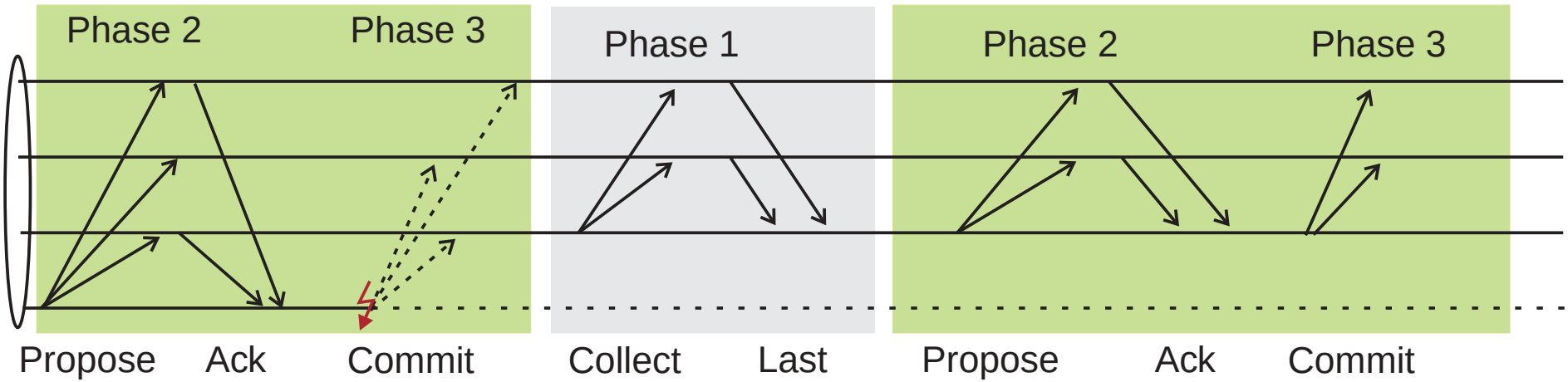
- Anführerbasierter Einigungsalgorithmus



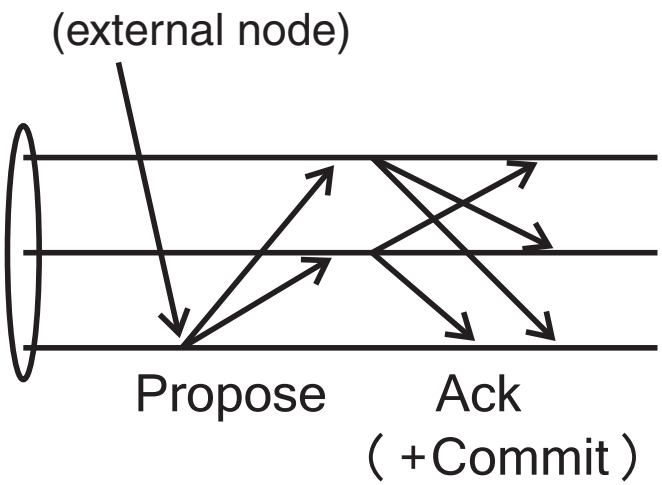
- Phase 2,3: Normalbetrieb (analog 2PC)
  Quorum (Mehrheit) der Knoten muss zustimmen
- Phase 1: Recovery nach Anführer-Ausfall
  Quorum (Mehrheit) der Knoten muss Zustandsinformationen liefern

Kommunikationsmuster: klassicher Paxos

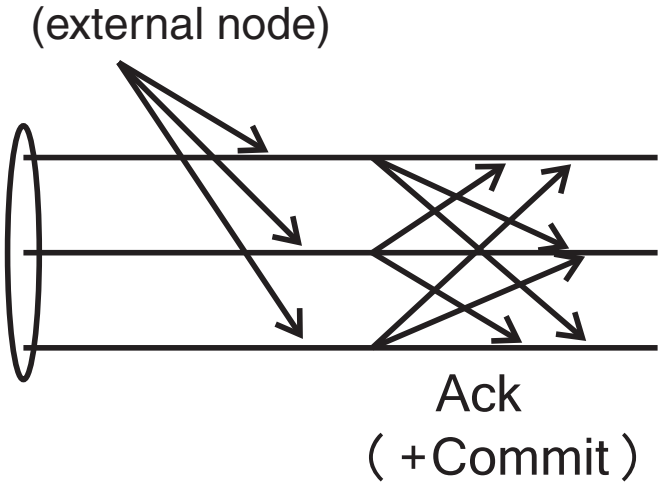
Anführerwechsel nach Crash



Kommunikationsmuster: schneller Paxos



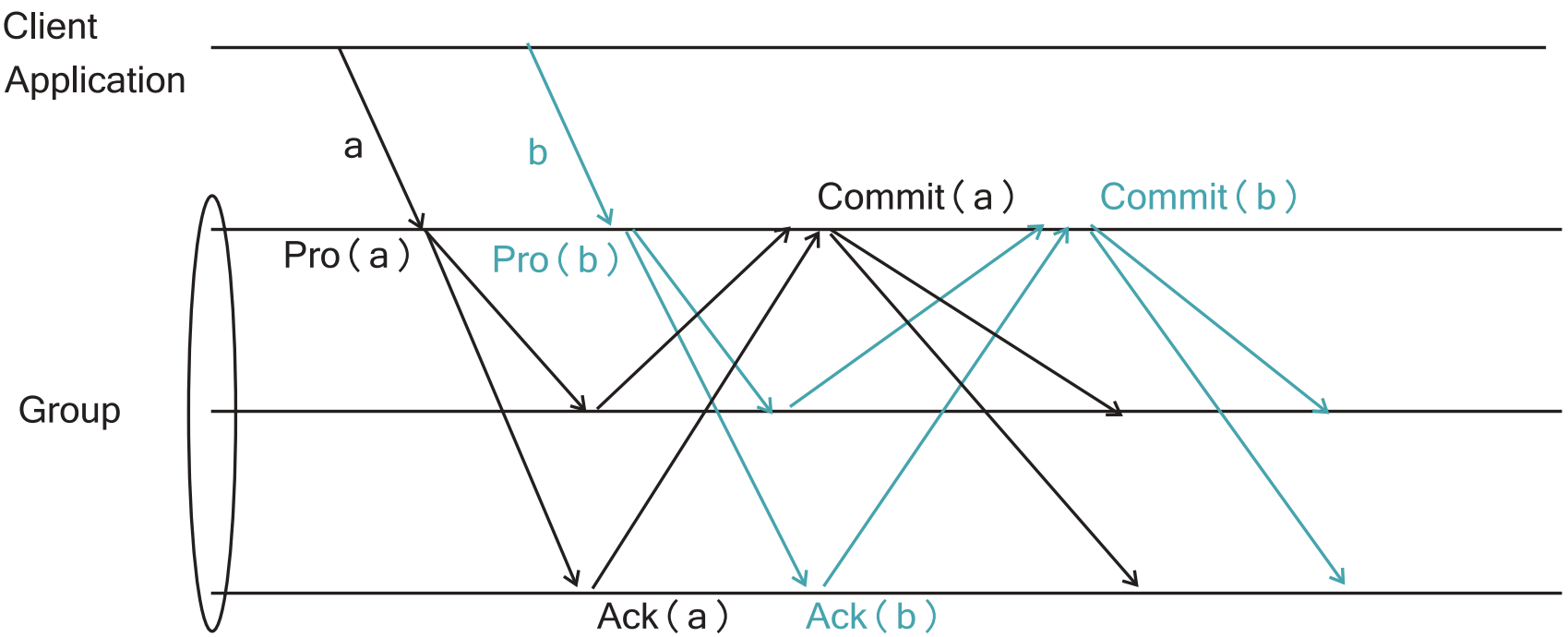
Fast Paxos



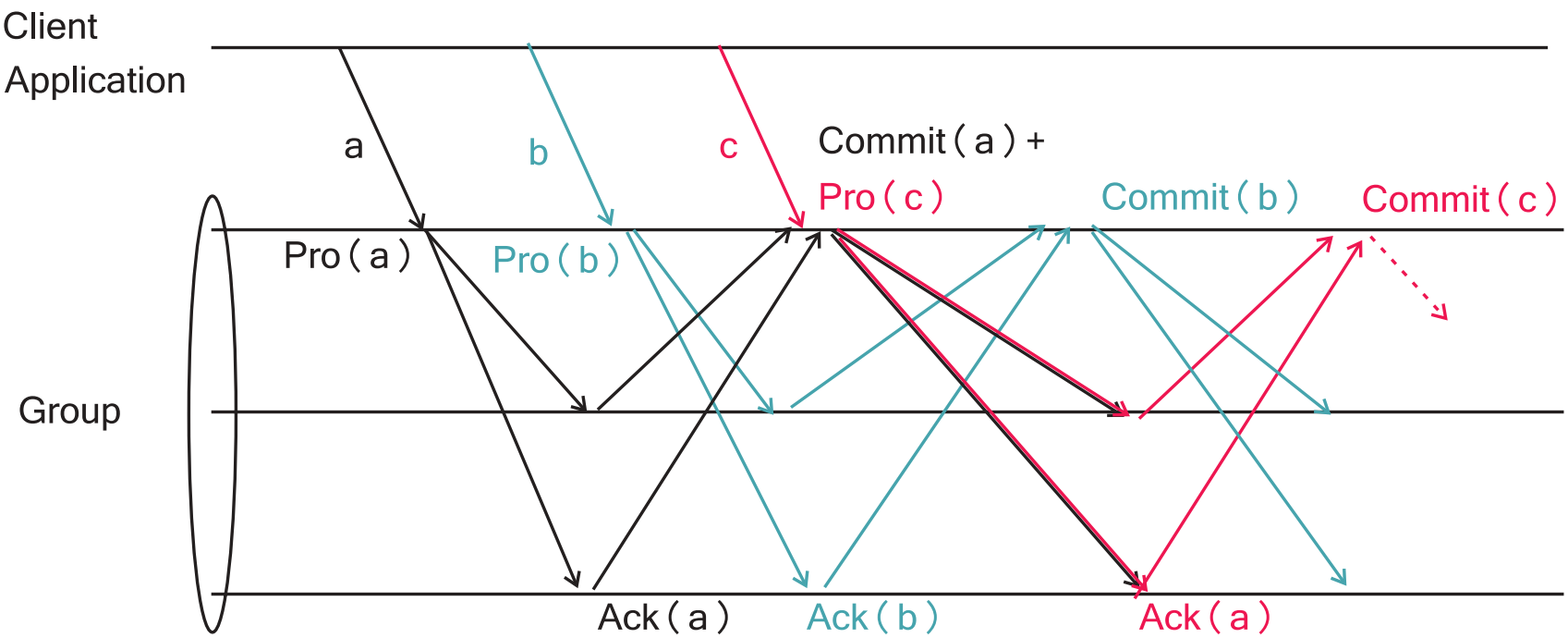
Ultra−Fast Paxos

- Fast Paxos: Optimierte Phase 2/3, 1 Delay Weniger, aber mehr Nachrichten
- Ultra Fast Paxos: 2 Delay weniger, optimistisch, Recovery nach Fehler

Parallelität von Entscheidungen

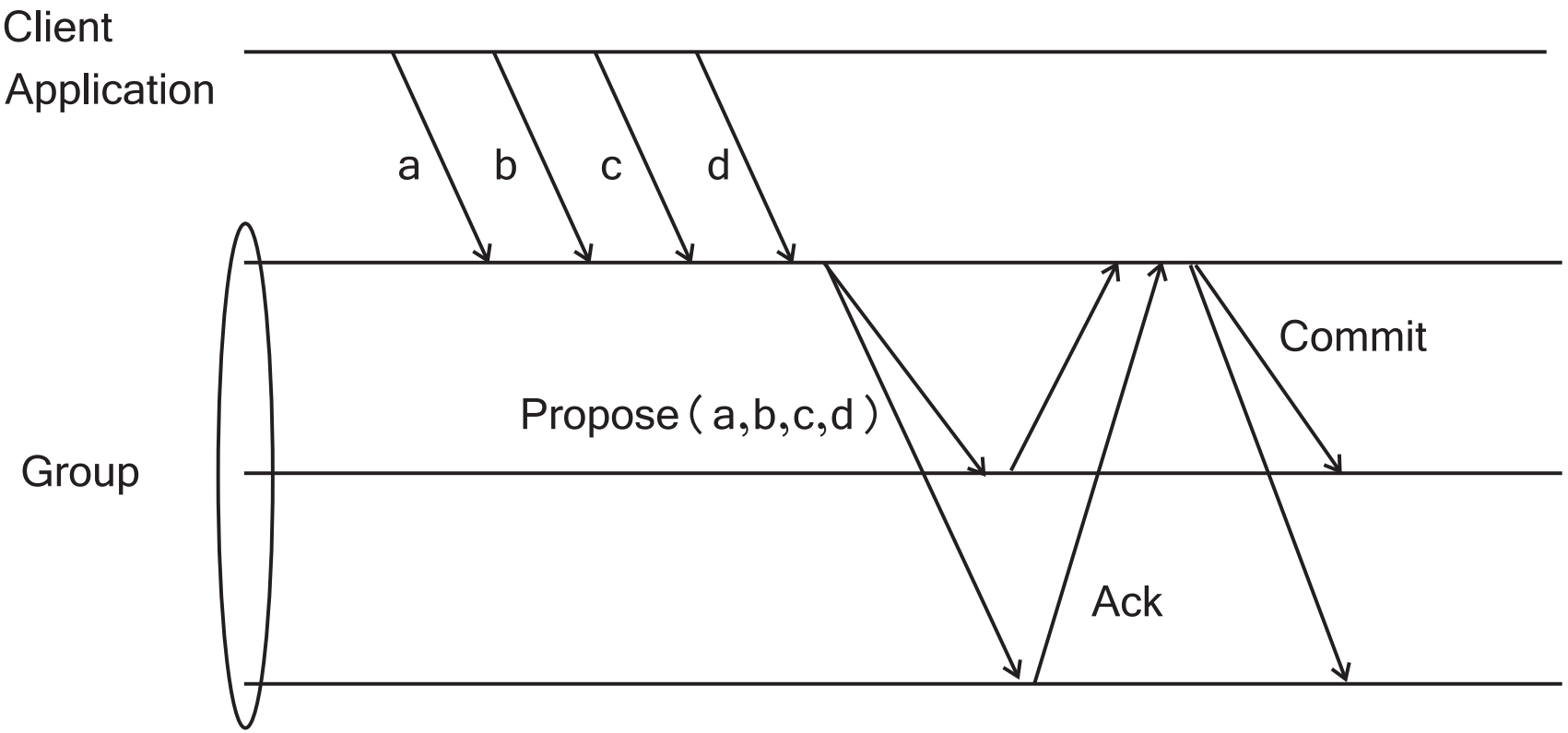


Parallelität von Entscheidungen



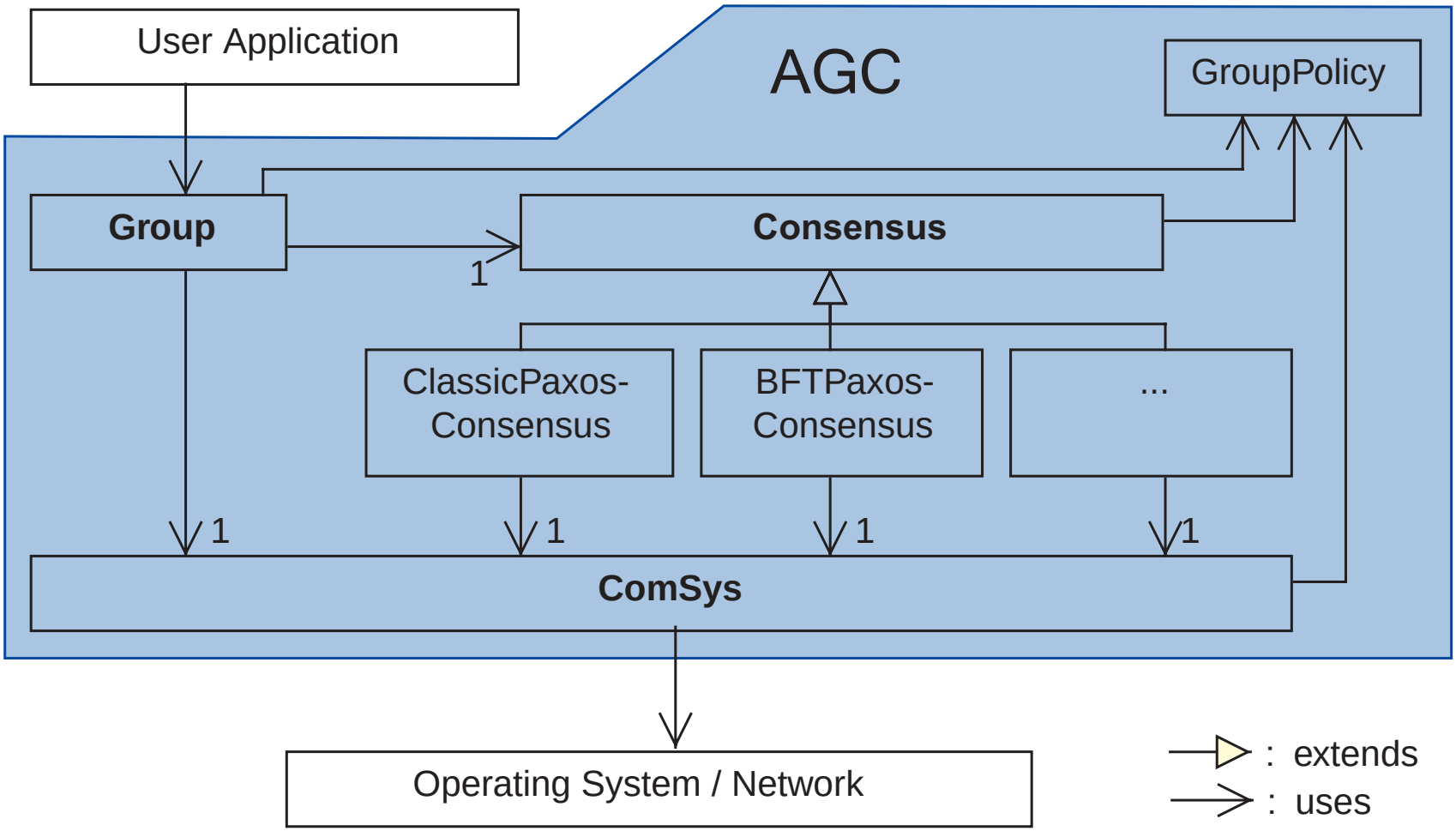
Sammlung von Nachrichten

Zustellung von N Nachrichten in einer Einigung



- Effizienzsteigerung (weniger Overhead)
- auf Kosten der Latenz (?)

Architektur



Was lässt sich rekonfigurieren?

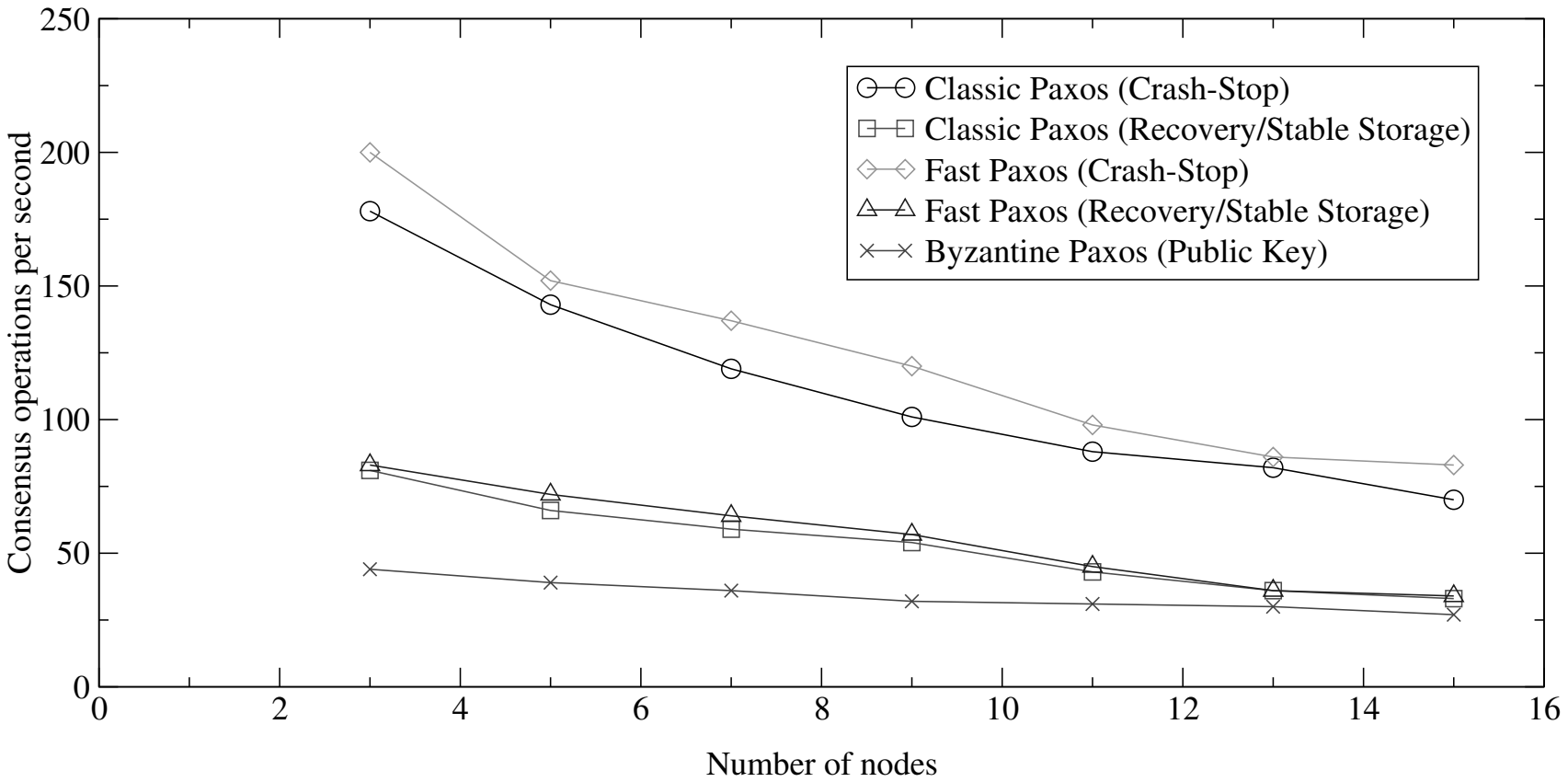
- Beteiligte Knoten (Join/Leave)
  - Join/Leave; Policy für erlaubte externe Knoten
- Einigungsvariante
  - Koordinierung bei Komplettaustausch erforderlich
  - Beschränkung auf N parallele Runden
- `ComSys`
  - TCP/IP, SSL, UDP, Multicast, ...
- Globale Gruppenparameter (z.B. Nachrichten/Wartezeit pro Einigung),
- Einigungsspezifische Parameter: Timing, Quorenmodelle

Evaluierung

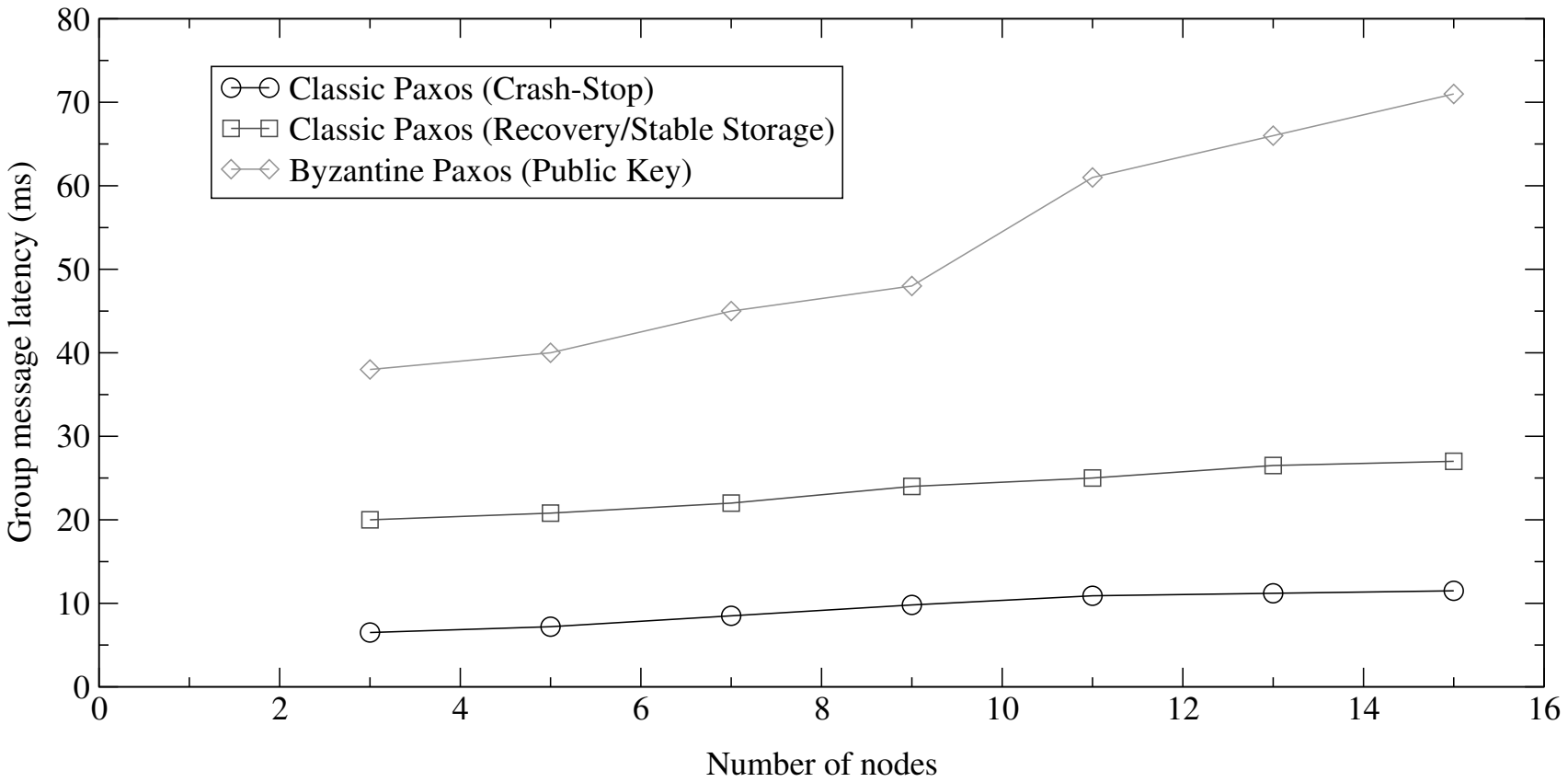
Untersuchung der Effizienz des Systems

- Verschiedene Varianten des Einigungsmoduls
- Variable Anzahl an Knoten
- Testumgebung mit 3...15 Rechnern
  - Lokales Ethernet (100Mbit/s), Linux-PCs (Pentium 4, 3.0 GHz)
- Untersuchung von:
  - Einigungsoperationen pro Sekunde
  - Latenz einer einzelnen Gruppennachricht

Einigungsoperationen pro Sekunde



Latenz einer einzelnen Nachricht



Zusammenfassung

Das AspectIX-Gruppenkommunikationssystem (AGC) bietet

- Effiziente, bedarfsgerechte Gruppenkommunikation
- Unterstützung verschiedener Fehlermodelle
- Dynamische Rekonfiguration zur Laufzeit