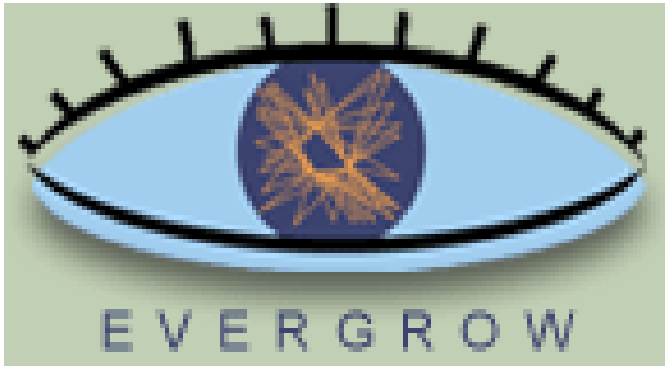
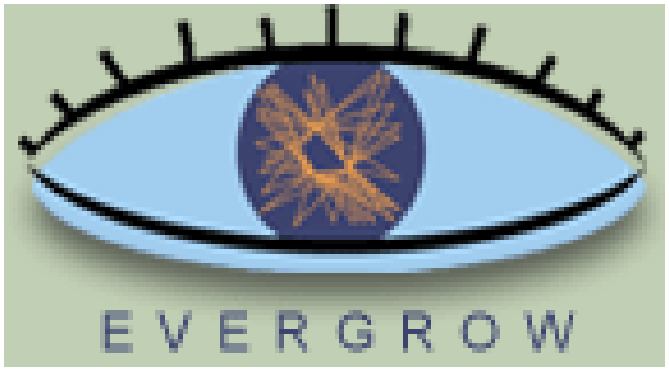


EVERGROW – Probing the Internet of 2025

Erik Aurell
SICS Swedish Institute of Computer Science
www.sics.se

LCSC, October 19, 2004



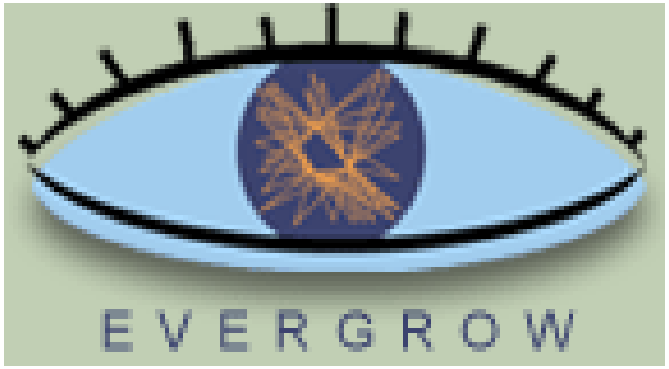
European Union 6th Framework Programme
Integrated Project EVERGROW (2004-2007)

ever-growing global scale-free networks, their provisioning, repair and unique functions

www.evergrow.org

Project management: SICS Swedish Institute of Computer Science
Scientific management: E.A. and Scott Kirkpatrick, HUJI, Israel

EVERGROW Partners



Research Institutes

1. SICS, Sweden
2. Abdus Salam ICTP, Italy
3. Collegium Budapest, Hungary
4. Istituto di Interscambio Scientifico, Italy
5. Central Laboratory of A.E.S., Egypt

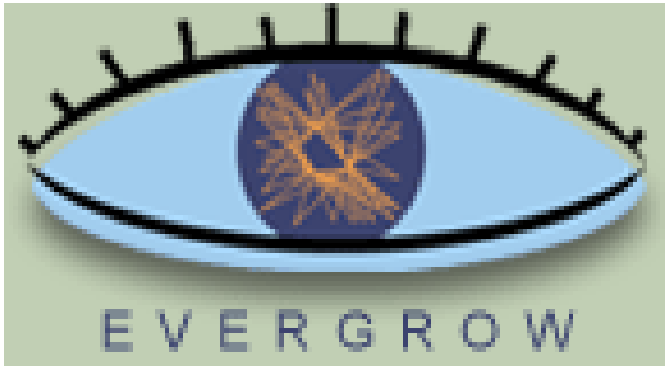
Companies

1. Ericsson AB, Sweden
2. France Télécom R&D, France
3. Sheer Networks, Israel
4. TeliaSonera, Sweden
5. Cetic , Belgium

Universities

1. Aston University, UK
2. École Nationale Supérieure, France
3. École Polytechnique Federale de Lausanne, Switzerland
4. Hebrew University of Jerusalem, Israel
5.KTH, Sweden
6.Köbenhavns Universitet, Denmark
7.Otto von Guericke-Universität, Germany
8.Technical University of Crete, Greece
9.Tel Aviv University, Israel
10.Université Catholique de Louvain, Belgium
11.Universidad Pública de Navarra, Spain
12.University of Oxford, UK
13.Université Paris-Sud 11 ”Orsay”, France
14.Istituto Nazionale per la Fisica della Materia, Italy
15.Universidad Rey Juan Carlos, Spain
16.CNRS, France

Goals



Means

distributed storage
ultimate RAID

distributed content delivery
ultimate AKAMAI

distributed information retrieval
ultimate GOOGLE

distributed storage and content delivery
ultimate GNUTELLA

Subproject 1
topology measurements
traffic measurements

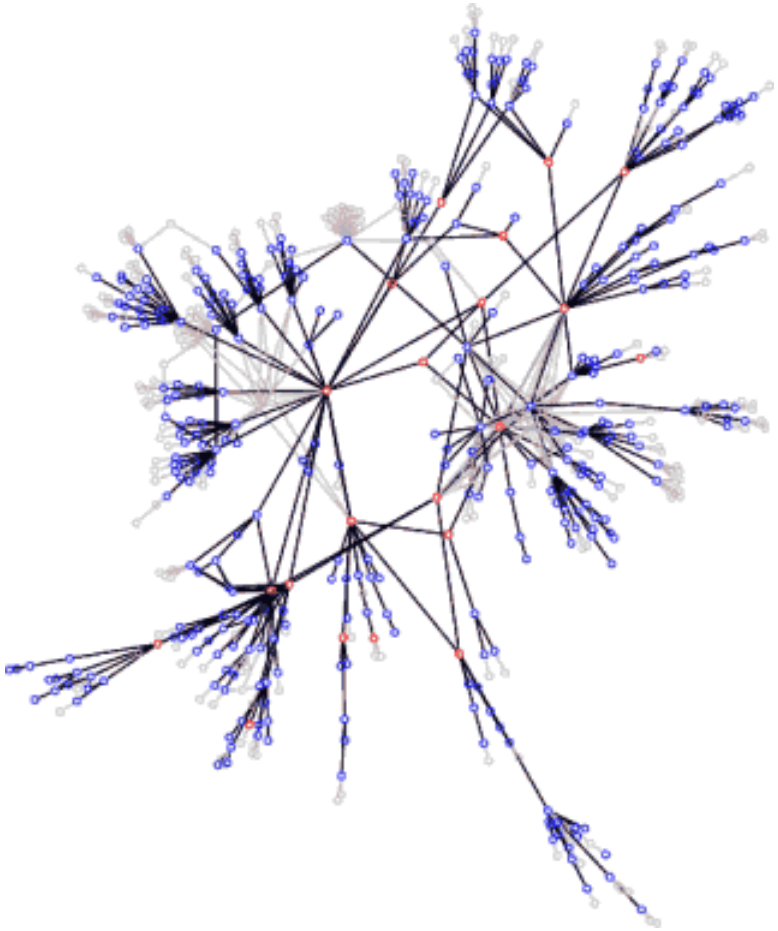
Subproject 2
distributed cluster

Subprojects 3-5
future P2P designs
message-passing
statistical physics
market mechanisms

Internet

a reminder

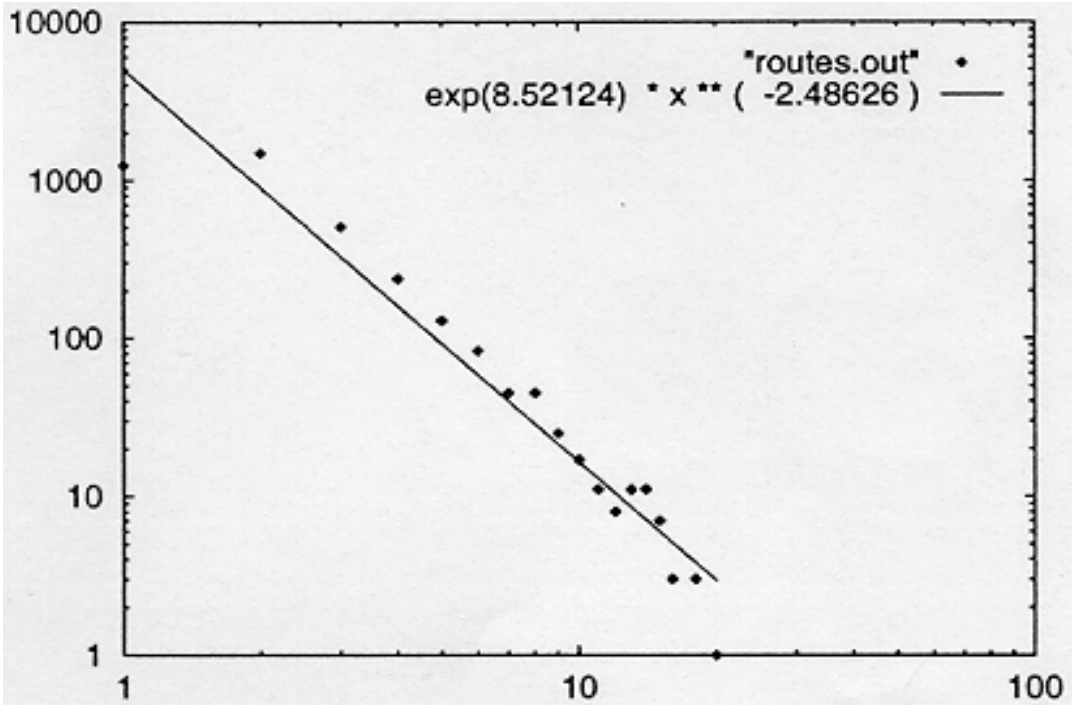
what is scale-free?



The Faloutsos Graph

1995 Internet router topology

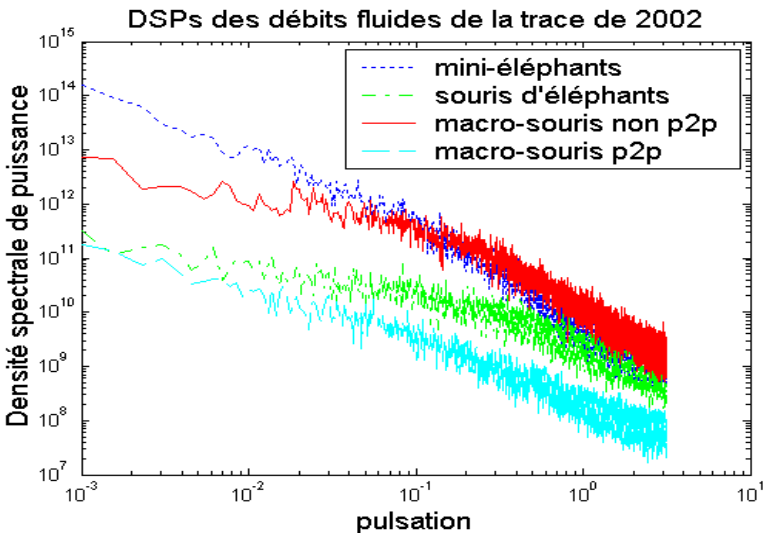
3888 nodes, 5012 edges, <k>=2.57



Overlay (Peer-to-Peer) networks on top of Internet

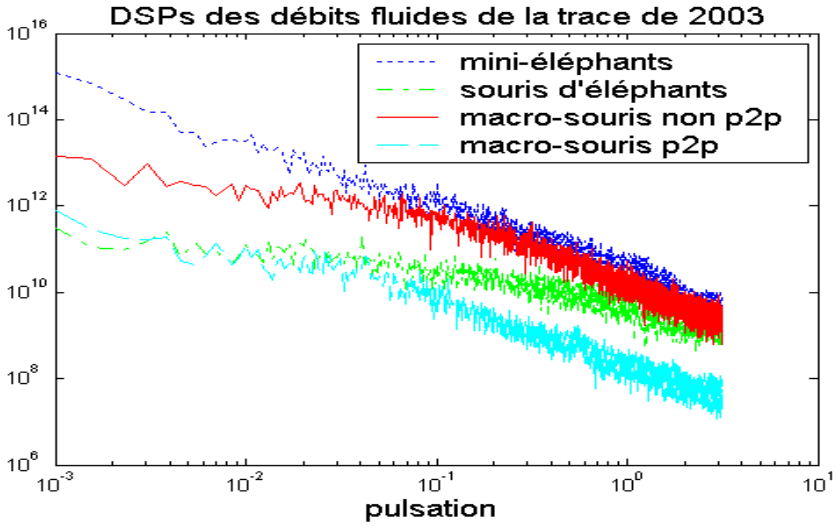
Account for more than three quarters of traffic on ISP level today

one day in 2002



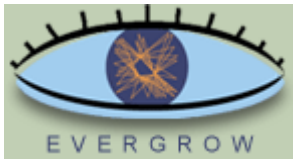
frequency

one day in 2003



frequency

Nadia Benazzouna and *Fabrice Guillemin*, France Telecom (2004)



Subproject 1
traffic and topology measurement programme

SW-based, freely down-loadable, measurement clients (DIMES), used primarily for topology discovery.

HW-based high-precision traffic measurement infrastructure of approx. 20 nodes, distributed in the European Internet

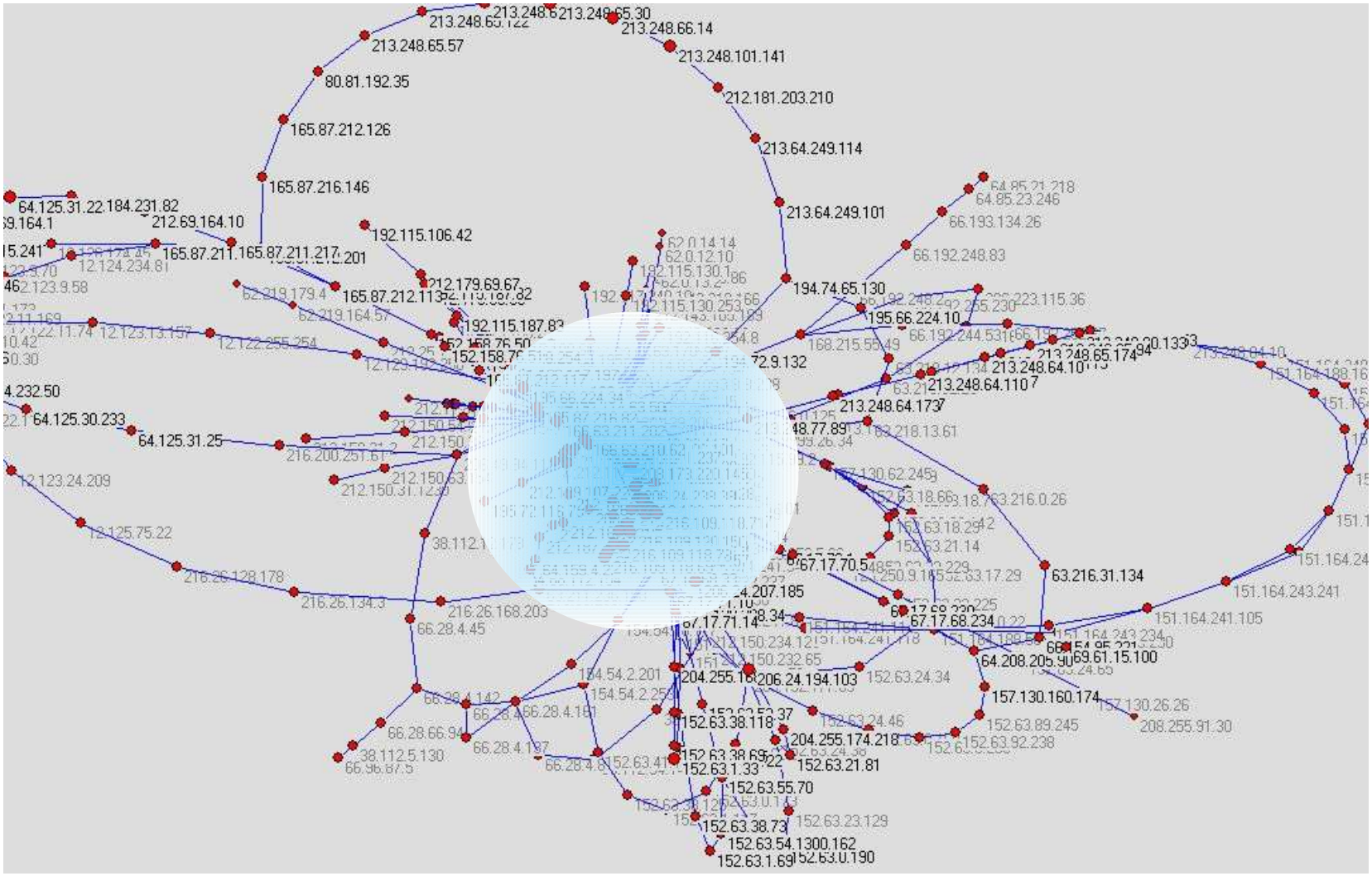
Making the measurement infrastructure publicly available for researchers (measurement management infrastructure)

Making large amounts of traffic measurement data publicly available (distributed data repository)

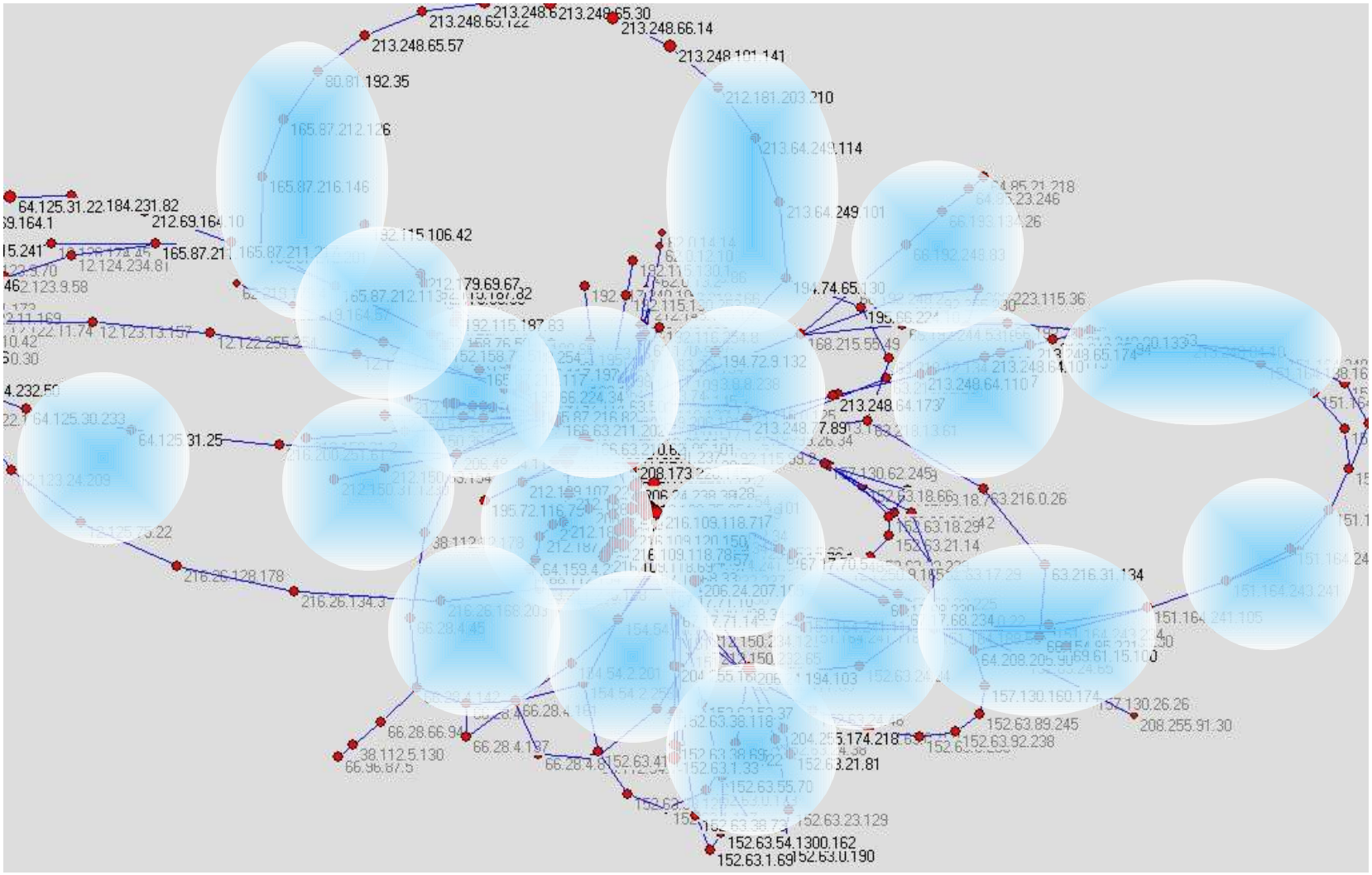
what is the problem with the Faloutsos graph? with past and present Internet mappings?

- Current projects are based on a few vantage points (e.g., Skitter at CAIDA).
  - Reveal mainly client-provider links
  - Miss many links that connect peering ISPs
  - The revealed topology tends to almost a tree
- Other problems:
  - Measurement rate depends on machine abilities: CPU, bandwidth
  - Easily noticed: “you ping me once a day, please stop”

measurements with few vantage points



measurements from many points



DIMES distributed measurements

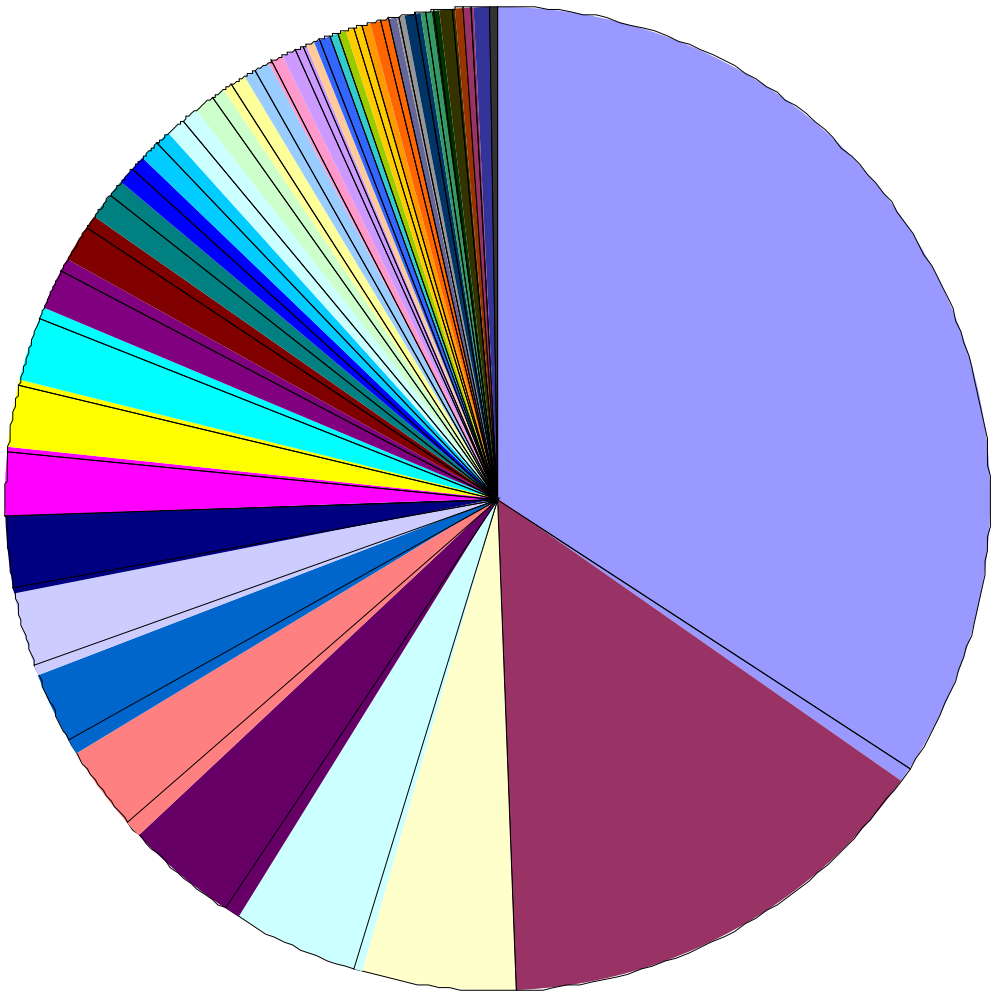
- Distributed measurement based on light agents
  - Ask the general public to download DIMES agent
  - Agents everywhere: no links are missed
  - Many agents: very low load on an agent
  - Stealth: same agent will not measure a target frequently
- Measurement types:
  - Current: connectivity, delay
  - Future: bandwidth, loss, etc.
- Aims: Generate maps, track internet dynamics over time

DIMES statistics (October 11th , 2004)

- Over 270 agents
  - Over 40 countries
  - All continents
  - ~60 are active daily
- Over 100,000 measurements a day
- Current map
  - Over 4000 ASes
  - About 10,500 AS links

twice as many links as Faloutsos....after one month!

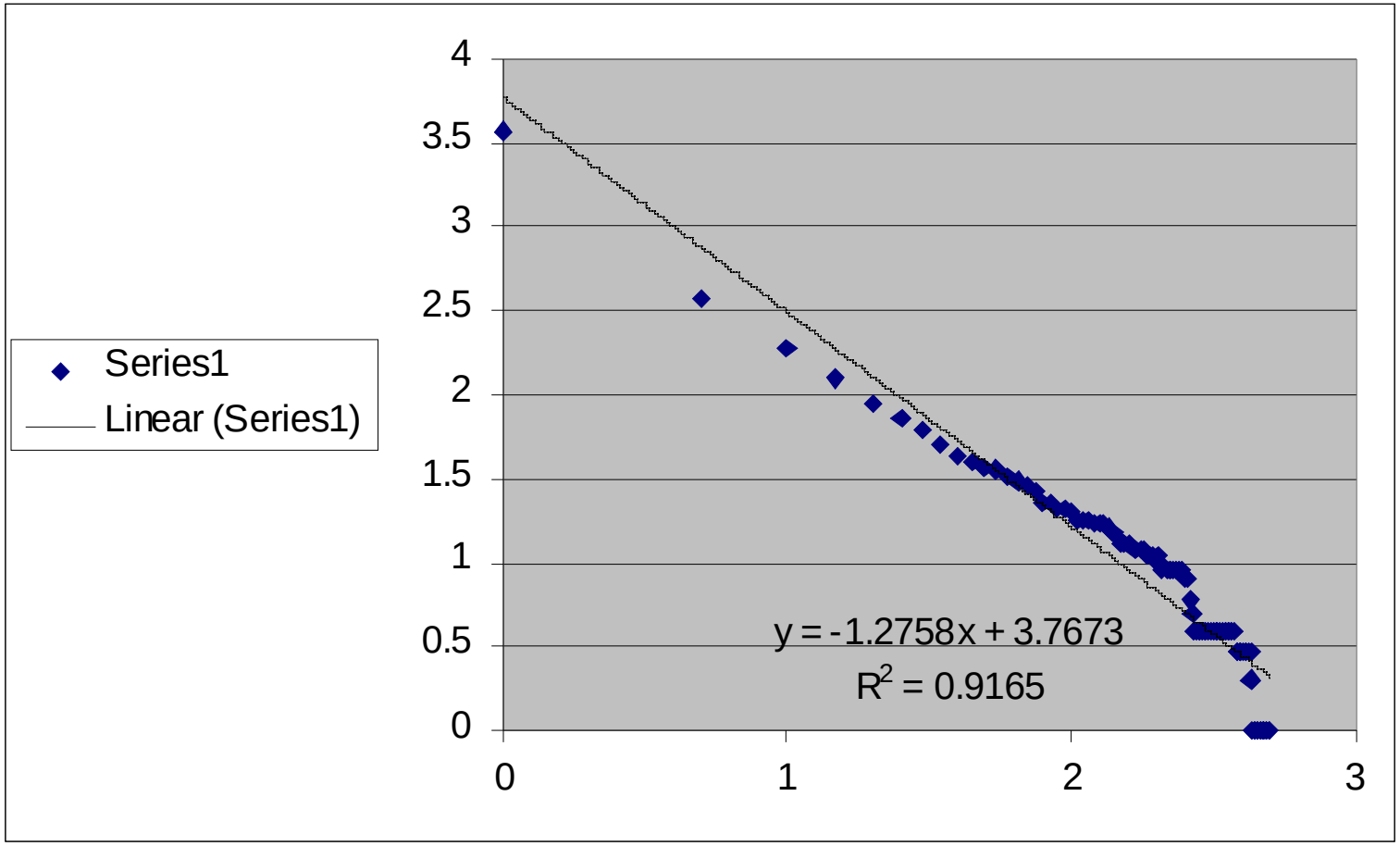
DIMES Agents by Country



- United States
- Israel
- no data
- Canada
- Sweden
- Spain
- Australia
- Denmark
- Germany
- France
- Hungary
- United Kingdom
- Netherlands
- Finland
- Korea (Republic of)
- Belgium
- China
- Japan
- Poland
- Italy
- Namibia
- New Zealand
- Norway

preliminary results are more fat-tailed than Faloutsos

AS degrees cumulative distribution



why measure?

- Benefit for network operators (enables more effective network management, including automatic control and optimisation of network resources)
- Benefit for end-users if we can make accurate tools available (enables more effective markets)
- Scientifically interesting to describe and try to understand the dynamics of Internet traffic

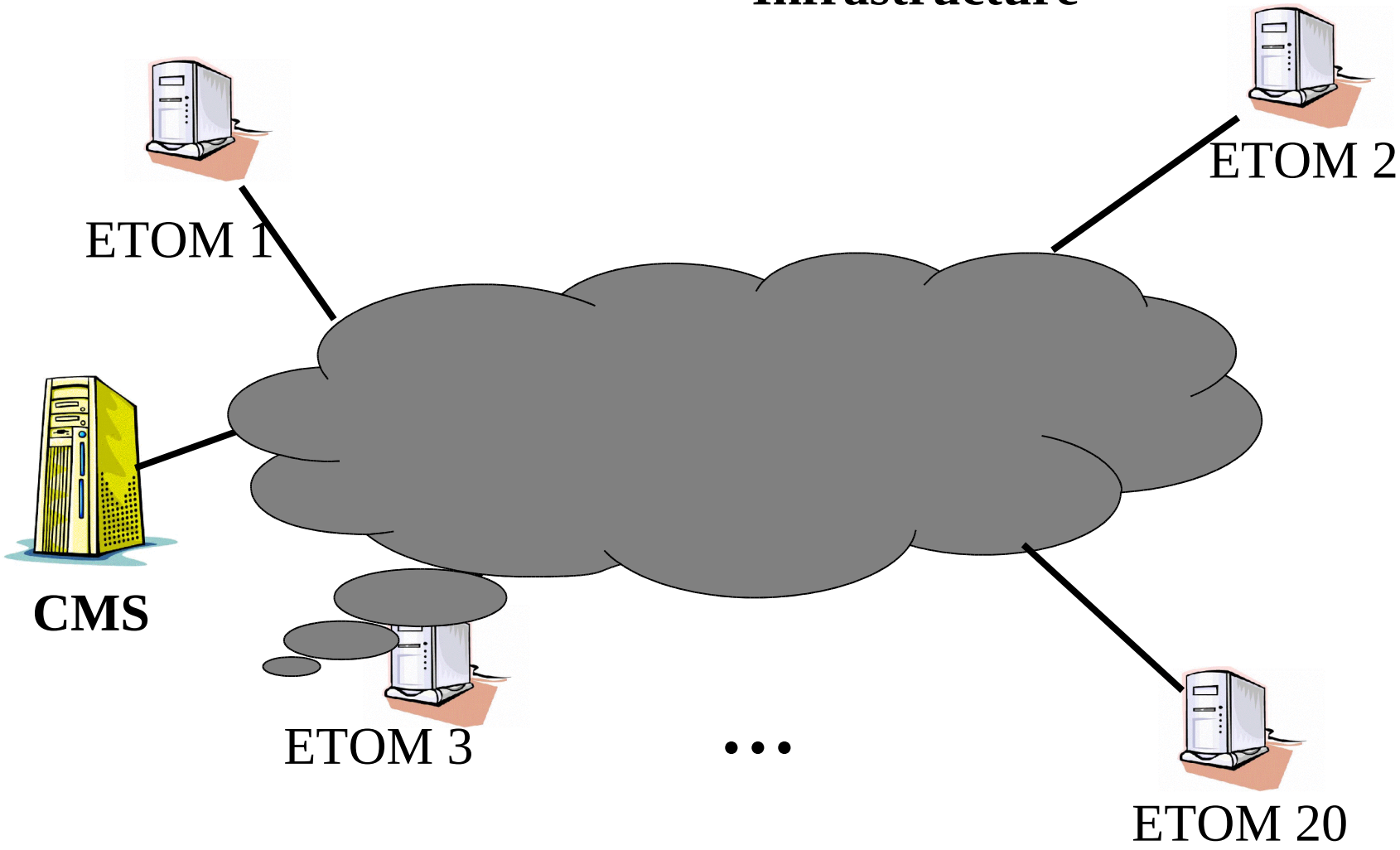
why active measurements?

- Enables measuring over network domains not under own administration
- It is the only way to measure end-to-end along a path

passive measurements will also be pursued in EVERGROW Partner France Telecom
necessary to test and evaluate active methods

Subproject 1
traffic measurements

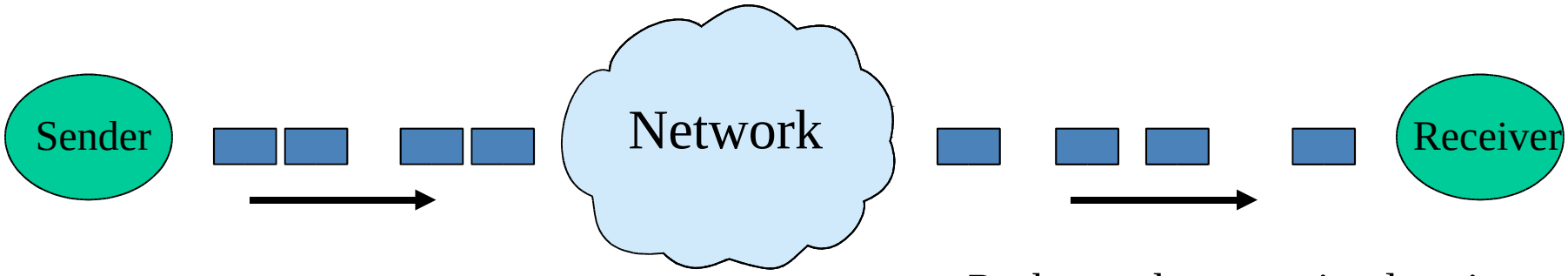
www.etomic.org
EVERGROW Traffic Observatory Measurement Infrastructure



Basic idea behind active measurements:

1. Inject “probe packets” into the network
2. Observe the effect exercised on the probes by the network
3. Draw conclusions about the state of the network

(step 3 is the hard one…)

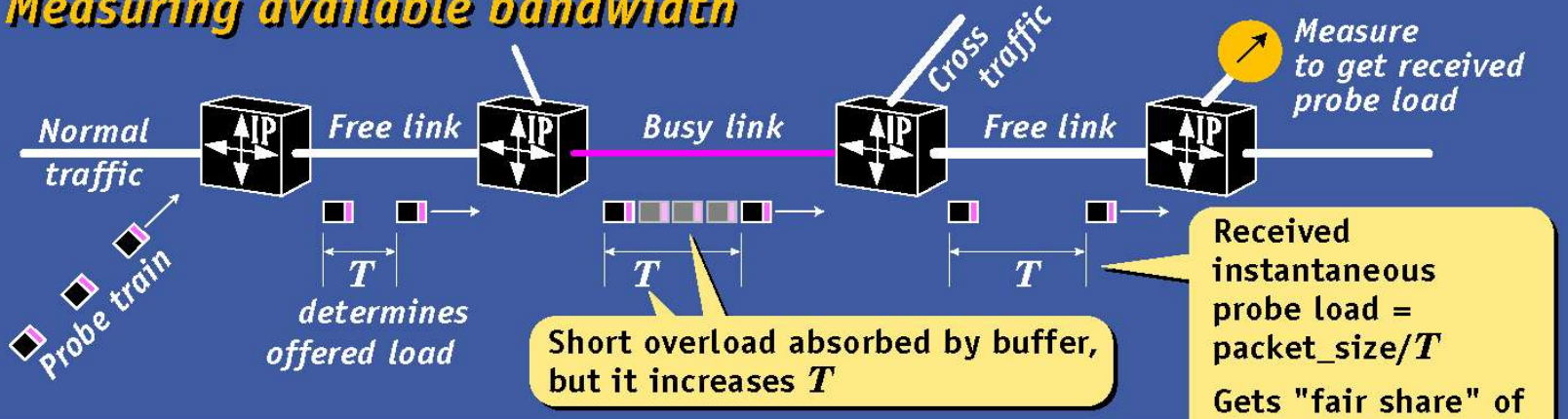


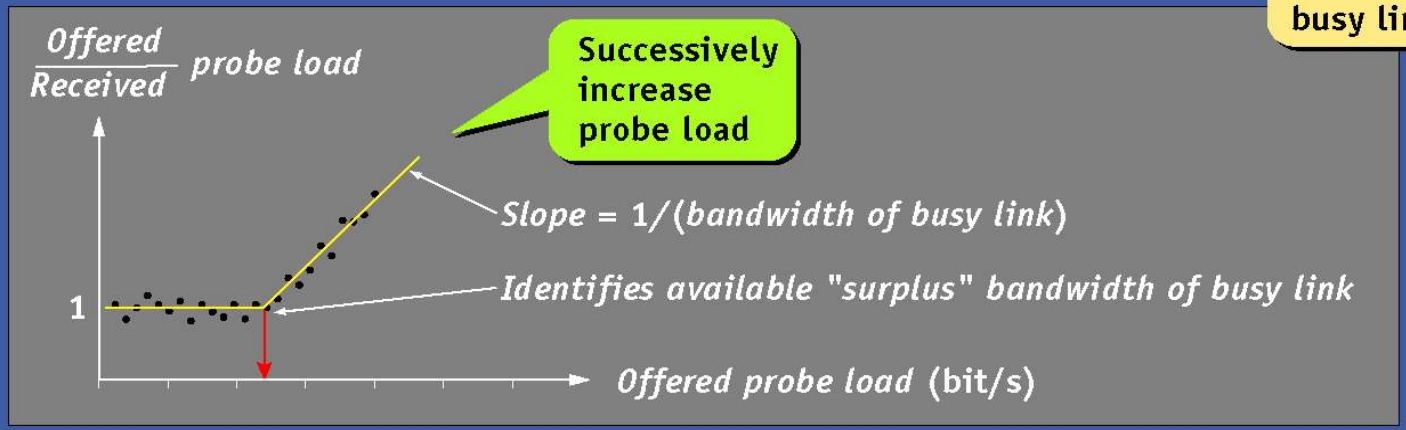
Probe packets transmitted at times
t_1, t_2, t_3, ...

Probe packets received at times
r_1, r_2, r_3, ...

Probe-Based Load Measurements

Measuring available bandwidth





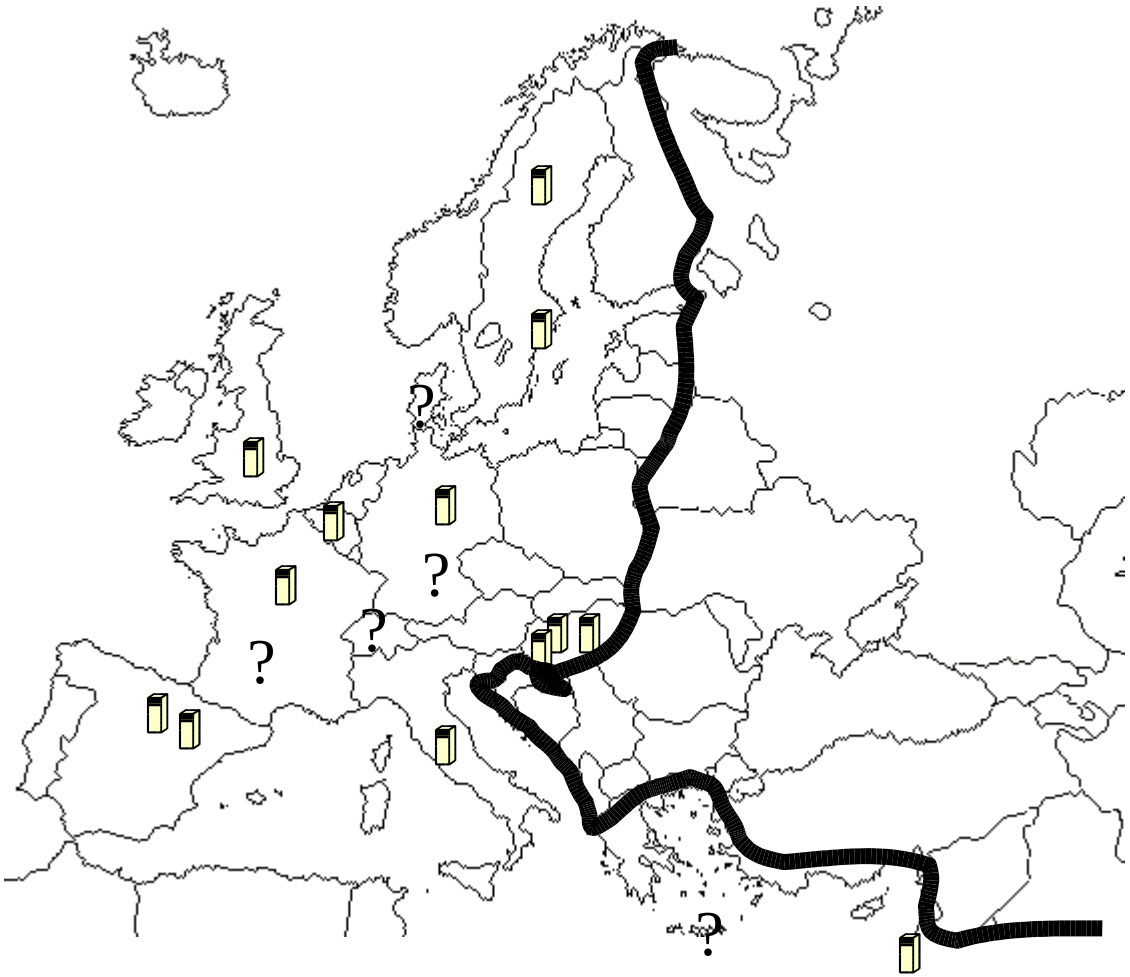
Subproject 1
traffic measurements

www.etomic.org
EVERGROW Traffic Observatory Measurement Infrastructure

what's new about this?

NLANR Active Measurement Project

the nodes will be in Europe

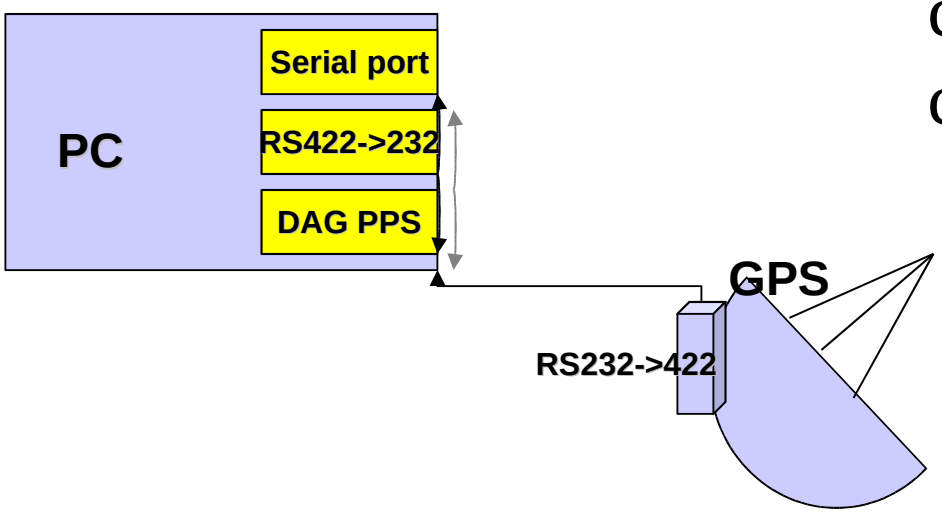


the Endace DAG 3.6GE card for active measurement probing the network

- PCI bus – 32 bit 33 Mhz
- Single port full packet capture at 10/100/1000 Mbit/s speed
- precise timestamping
- Burst of patterned traffic generator – sending special packets at 10/100/1000 Mbit/s – designed for us

time synchronization - the GPS module

- Garmin 35HVS GPS reciver
- 1 us PPS signal
- RS 232 – RS 422 converter – max 100m cable



Objectives

- Support three computing activities for EVERGROW
  - Virtual Observatory to hold EVERGROW Web measurements
    - SP1
  - Computational platform for study of message-passing and selfish optimization – traditional distributed computing issues
    - SP4,5
  - Simulate and eventually deploy novel Peer-2-Peer services
    - SP3

- First year – do useful work for all sites (“pre-GRID”)
- Second year – add GRID building blocks, esp for SP3

Process followed

- EVERGROW Executive negotiated with suppliers for basic configuration, price, before project started
- EVERGROW Consortium accepted configuration, 8 sites funded to obtain this hardware from local or common suppliers (time-consuming process)
- Basic configuration:
  - IBM blade server with 14 dual P4 blades, each 4 GB memory, dual Gb Ethernet switches, redundant power and management
  - Dual IDE 40GB hard files in each blade.
  - Front end 2u server for management and external access, also dual P4, 4 GB
  - Storage server with .8 TB SCSI storage, also dual P4, 4GB.

System structure

- GPFS manages two local volumes: in blade, in external servers. Each GPFS node can see the other clusters.
- VPN connects all blades regardless of site.
- Common release of kernel, file structure, and consistent naming, addressing conventions managed using CSM tools.
- Each user can log in to one site's management server. That sites sysadmin is responsible for validating users, and putting any constraints on access. Systems admins and selected power users can log in to all sites.

Status (Oct 2004)

- Four sites operational, eight expected online by year end
- Two sites setup VPN, two sites demonstrate GPFS, two sites responsible for batch scheduler selection and initial tuning. One site taking the lead in access control and naming.
- Later four sites use recipes from the first four.

- Ask us again at year-end…

deployed P2P systems are random overlays

robust to node failure
completely decentralized

no guarantees
difficult to find rare things
searching by broadcast
huge searching overheads
Gnutella, *KaZaa* etc.

structured overlays, or **D**istributed **H**ash **T**ables

time guarantees
also completely decentralized
search by key look-up

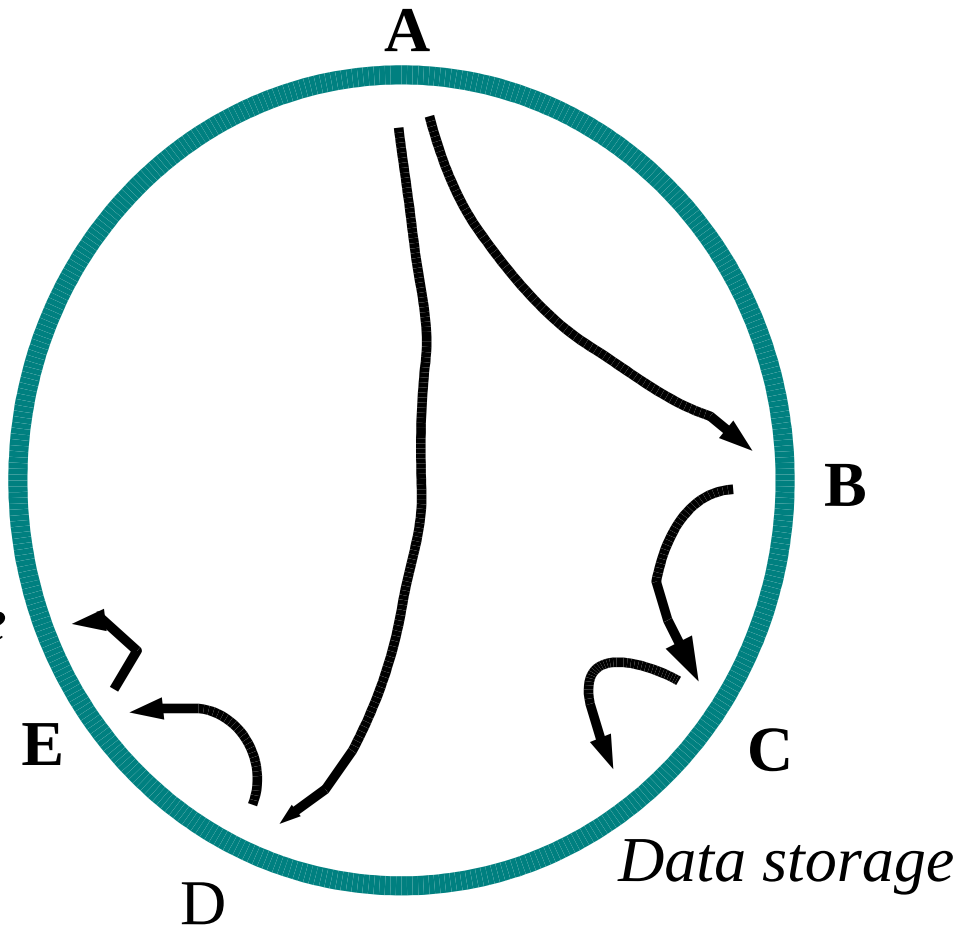
needs stabilization
robustness under churn?
Chord, *Pastry*, *Tapestry*, *DKS*
Skipnet, *Koorde*, *Kademlia*,
ViceRoy, etc.

Shown by *Rhea et al* (Berkeley-Intel collaboration) [USENIX2004]
that Chord and Pastry do not work under observed KaZaa churn

Churn = nodes join and fail all the time

structured overlays (**DHT**s) under churn are best analyzed as non-equilibrium physical systems

Chord
data keys and nodes on a ring
nodes store data for keys
nodes keep track of fingers
$1/2\log(P)$ hops in look-up



when nodes die fingers are lost
stabilization necessary

dimensional analysis of DHTs (Chord) under churn

$\lambda_s = 1/\tau$ rate of stabilization, per node and time

λ_J rate of new nodes joining, per node and time

λ_F rate of nodes failing, per node and time

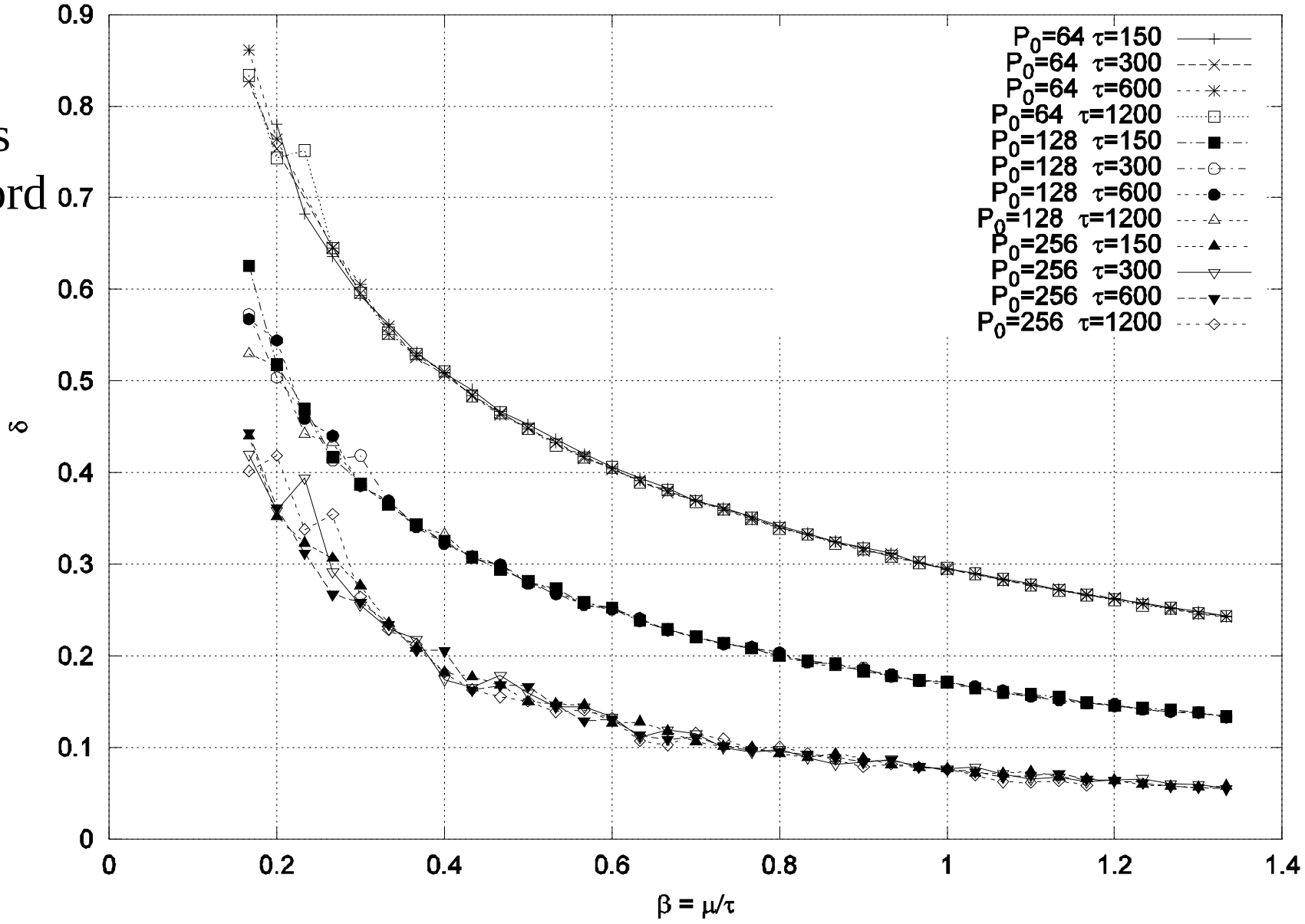
$\lambda_J = \lambda_F$ in steady state, rates of failures and joins balance

$\gamma = \lambda_s / \lambda_F$ dimension-less ratio of stabilization and failure

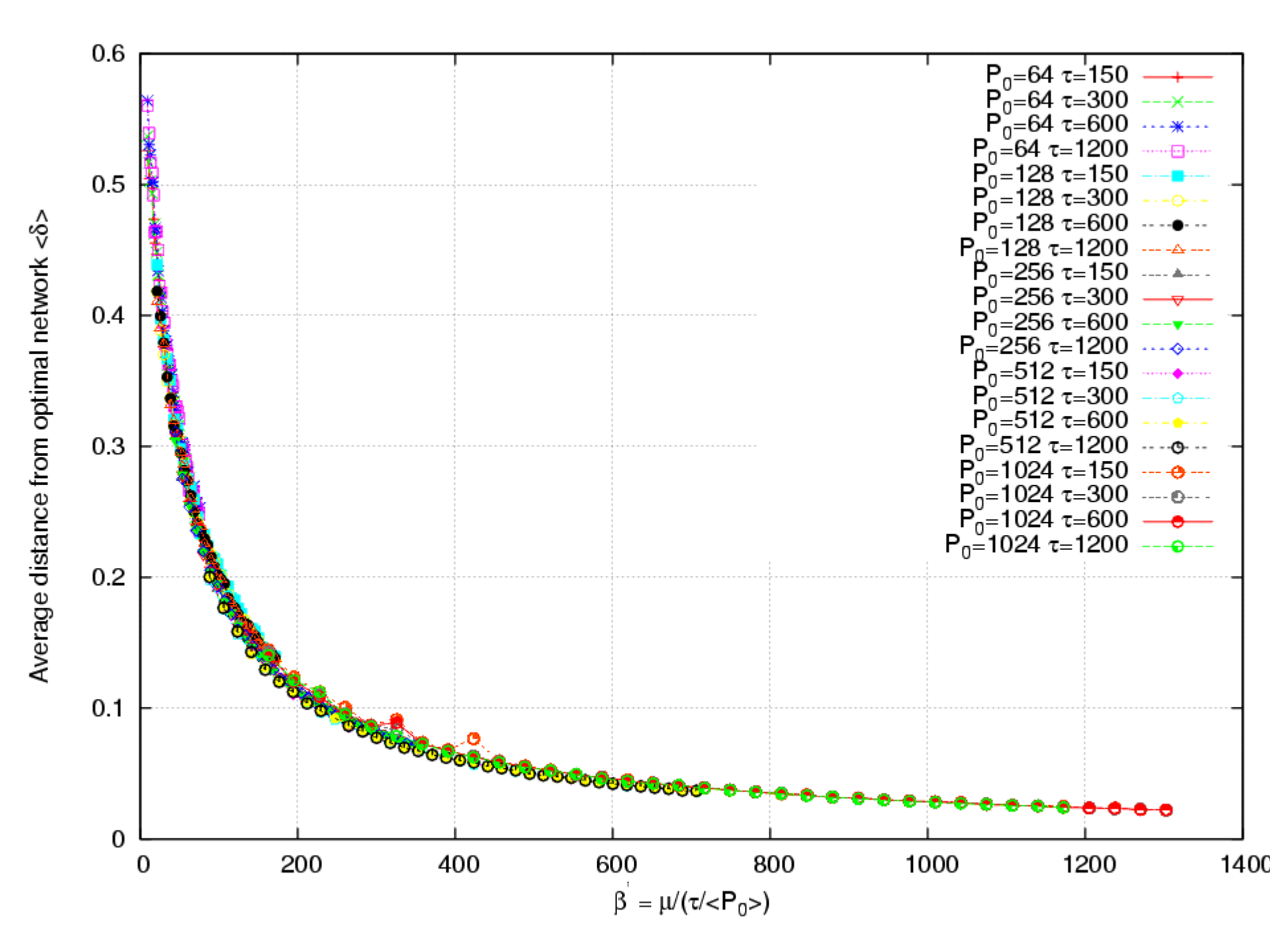
μ rate of new nodes joining the full network

$\mu = \mathrm{P}\,\gamma$ customary, though not intensive, variable

wrong fingers in Chord under churn



S. El-Ansary, E. Aurell, P. Brand, S. Haridi, LNCS (2004)


Average distance from optimal network <δ>
0.6
0.5
0.4
0.3
0.2
0.1
0
0
200
400
600
800
1000
1200
1400
β' = μ/(τ/<P0>)
P0=64 τ=150
P0=64 τ=300
P0=64 τ=600
P0=64 τ=1200
P0=128 τ=150
P0=128 τ=300
P0=128 τ=600
P0=128 τ=1200
P0=256 τ=150
P0=256 τ=300
P0=256 τ=600
P0=256 τ=1200
P0=512 τ=150
P0=512 τ=300
P0=512 τ=600
P0=512 τ=1200
P0=1024 τ=150
P0=1024 τ=300
P0=1024 τ=600
P0=1024 τ=1200

master equations of Chord under churn

simple in principle, laborious in practice
here is a sketch of the analysis of internode distances
to compare fractions of failed fingers of look-ups quantitatively
with simulations, you need to keep many system details
note these are results for **one** version of Chord
there at least four different (published) versions of Chord

$$N_m(t+dt) = N_m(t) - 1 \quad \text{with prob} \quad (2\lambda_F dt) N_m / P$$

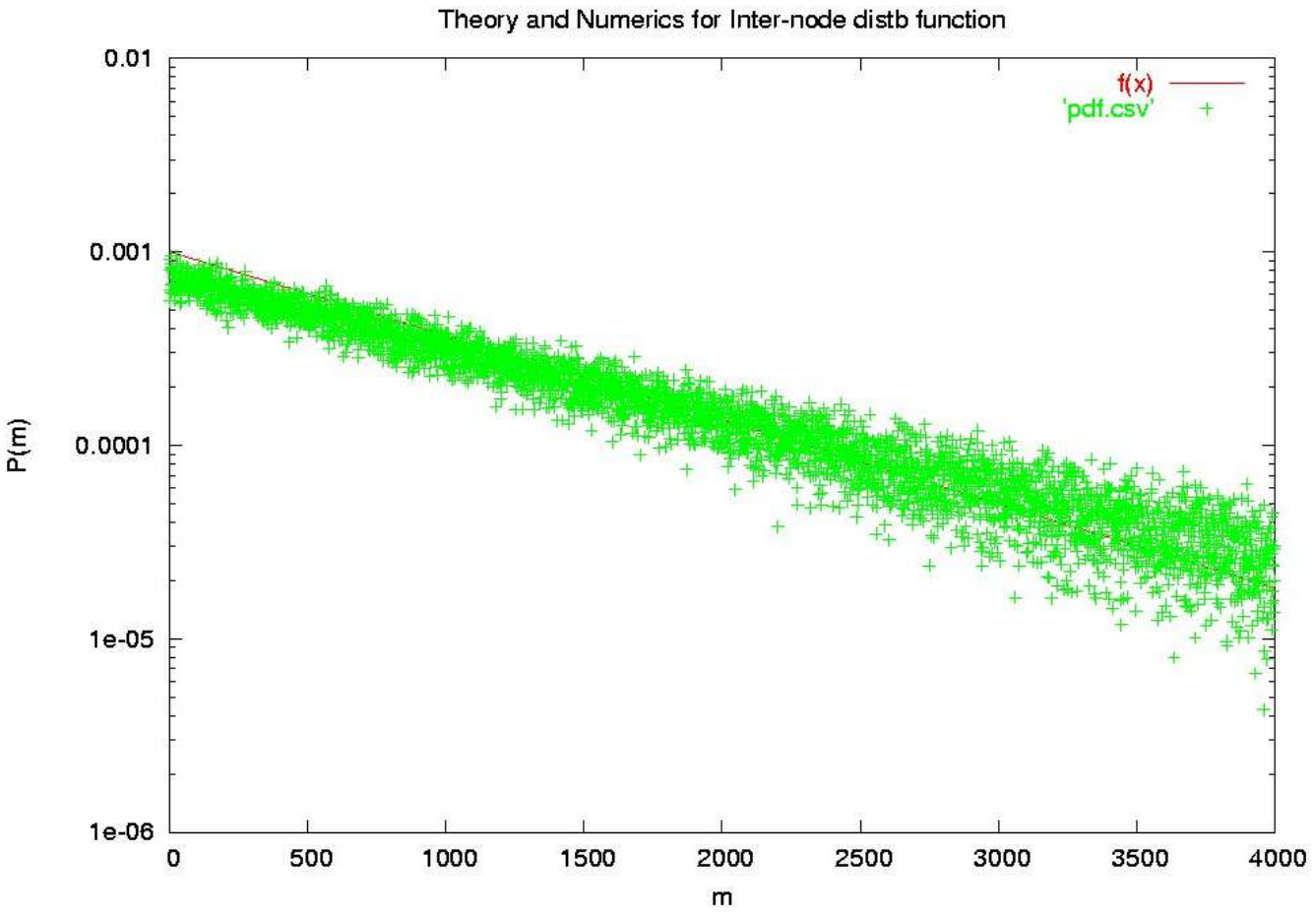
$$N_m(t+dt) = N_m(t) - 1 \quad \text{- " -} \quad \frac{(P\lambda_J dt)}{N-P}(m-1)\frac{N_m}{P}$$

$$N_m(t+dt) = N_m(t) + 1 \quad \text{- " -} \quad \frac{2N_{m'}(\lambda_F dt)}{P}\frac{N_{m-m'}}{P}$$

$$N_m(t+dt) = N_m(t) + 1 \quad \text{- " -} \quad (\lambda_J dt)\frac{2}{N-P}\sum\frac{N_{m'}}{P}$$

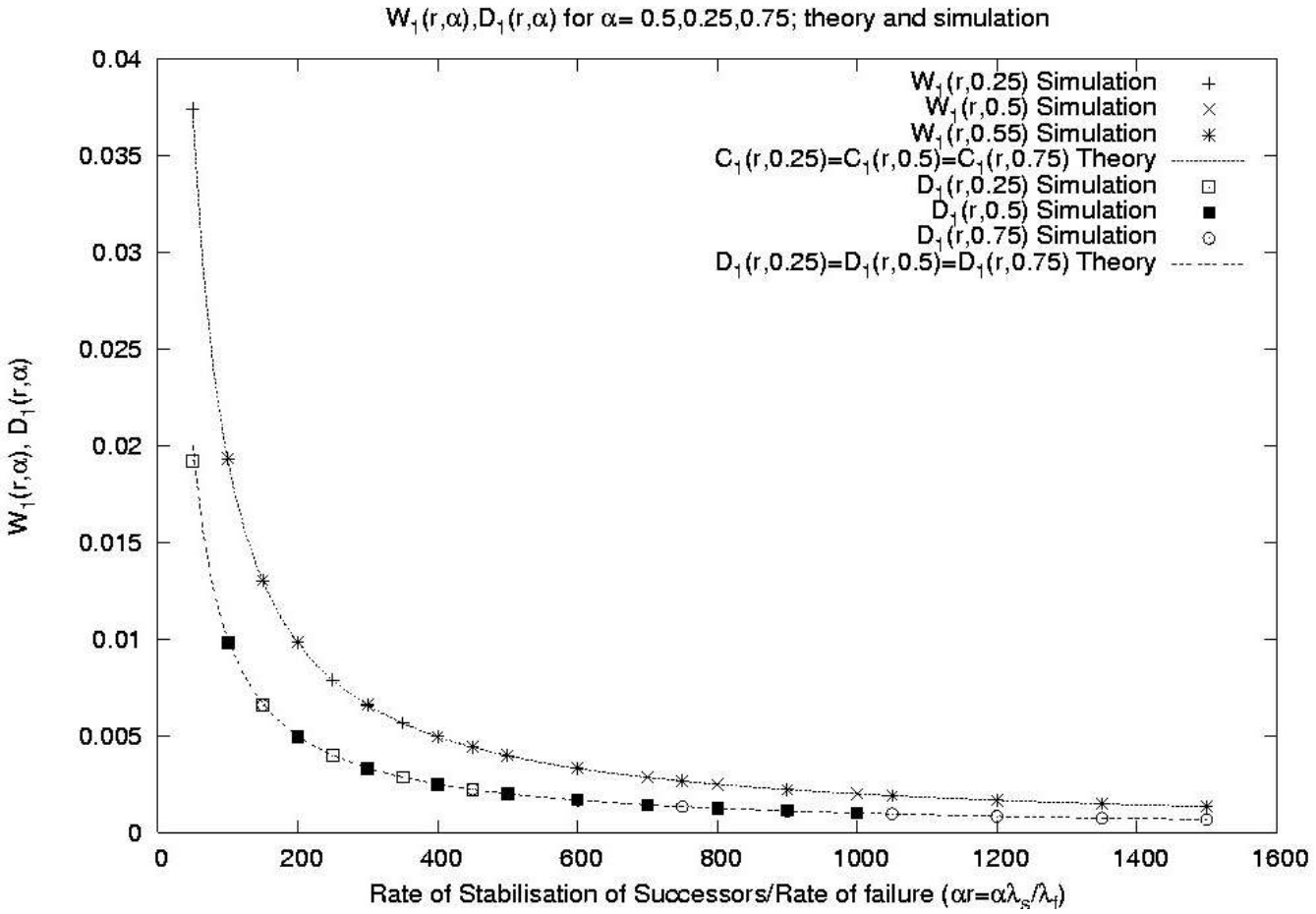
solution of master equation for internode distances is obvious

$$N_m = P\,\rho^{(m-1)}(1-\rho)$$



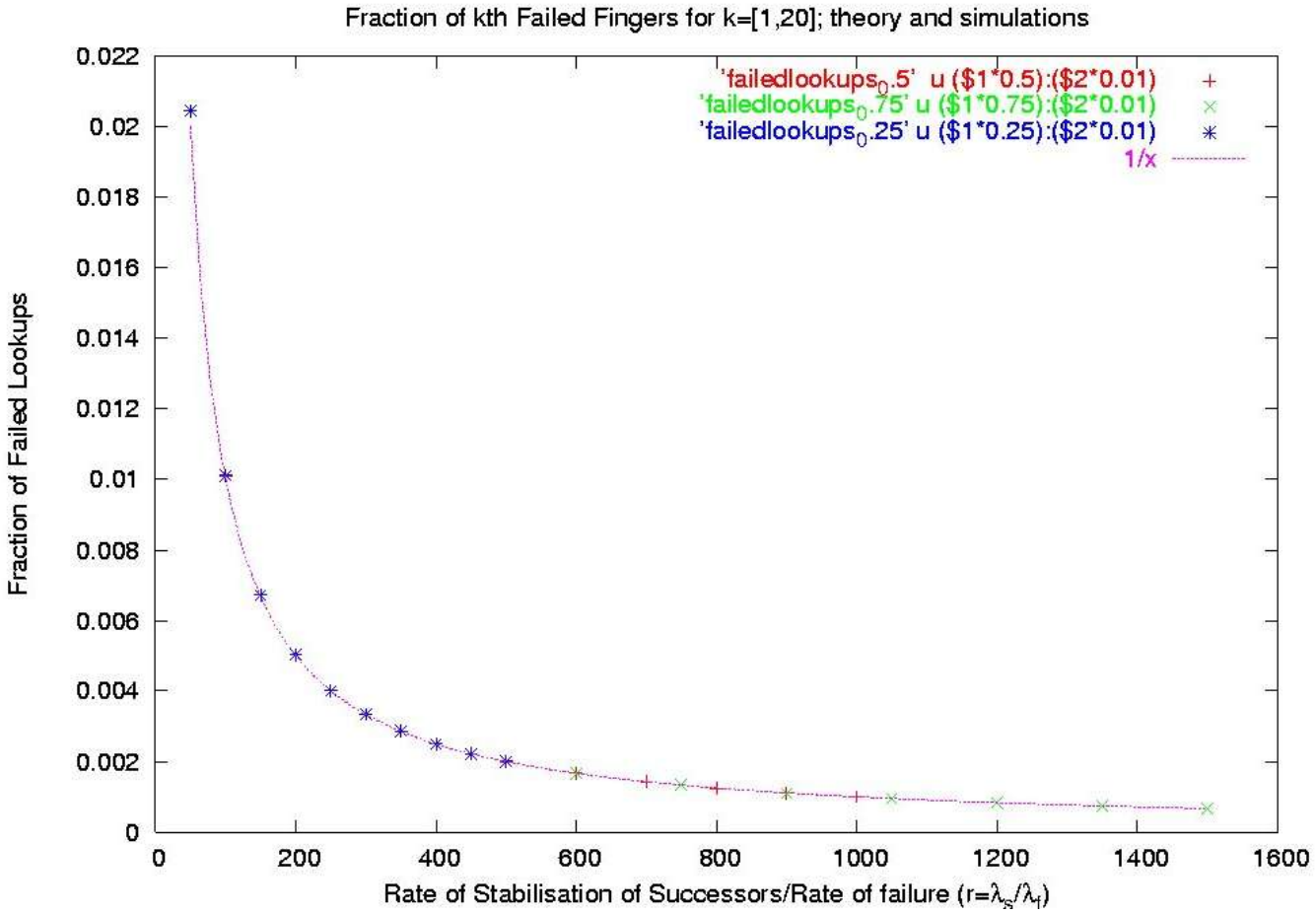
S. El-Ansary, S. Krishnamurthy, E. Aurell, S. Haridi, (2004)

probability of failed and wrong fingers of shortest length



S. El-Ansary, *S. Krishnamurthy*, *E. Aurell*, *S. Haridi*, (2004)

fraction of failed look-ups



S. El-Ansary, S. Krishnamurthy, E. Aurell, S. Haridi, (2004)

conclusions on analysis of future P2P systems

simulations require huge computational resources

emulations will be even worse

it helps to think before during and after you compute

simple physical arguments are useful to understand P2P systems

systems with detailed public specifications can be analyzed
in detail using master equations

large structured overlays have a hard time coping with churn

there is lots to do still!

European Union 6th Framework Programme
Integrated Project EVERGROW (2004-2007)

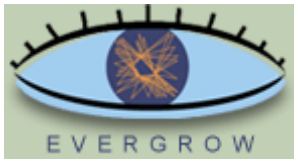
ever-growing global scale-free networks, their provisioning, repair and unique functions

www.evergrow.org

***Information Society and its Technologies* (IST) Directorate**

***Future and Emerging Technologies* (FET)**

Proactive Initiative **"Complex Systems Research"**



special thanks to

Scott Kirkpatrick

OTHER EVERGROW FRIENDS

Gabor Vattay
Yuval Shavitt
Svante Ekelin
Anders Rockstrom
Fabrice Guillemin

EVERGROW TEAM, SICS

Supriya Krishnamurthy
Sameh El-Ansary
Seif Haridi
Per Brand
Luc Onana Alima
Ali Ghodsi
Kersti Hedman
Janusz Launberg