Distributed Computer Systems = Making Computer Systems Scalable, Reliable, Performant, etc., Yet Able to Form an Efficient Ecosystem

Co-sponsored by:

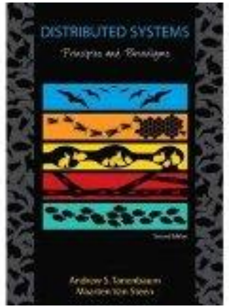
@AIosup

Prof. dr. ir. Alexandru Iosup

What is a Distributed System?

"You know you have a distributed system when the crash of a computer you've never heard of stops you from getting any work done."
- **Leslie Lamport** in Security Engineering, Ch.6

"A collection of independent computers that appears to its users as a single coherent system - **Steen and Tanenbaum** in Distributed Systems: Principles and Paradigms, 2nd Edition, 2006



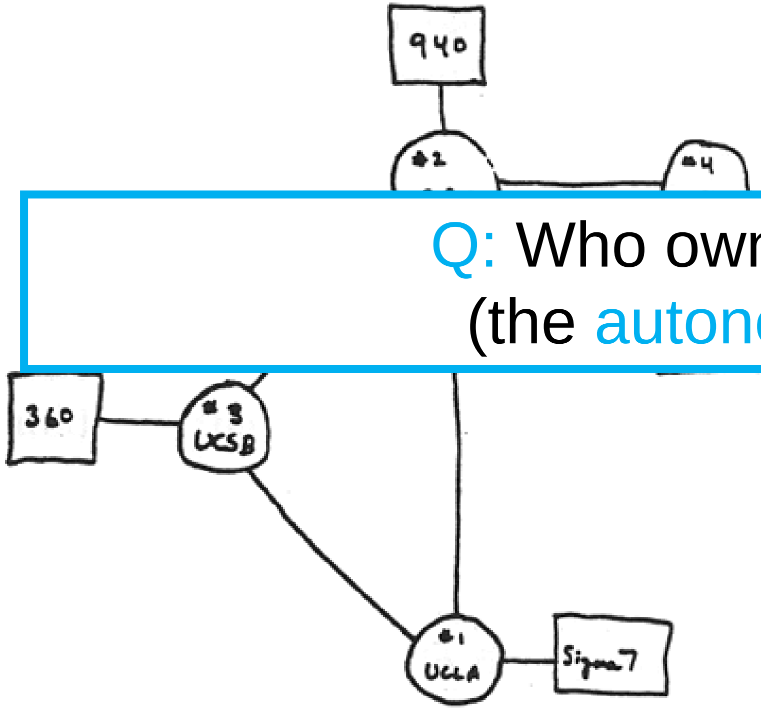
"A collection of ***autonomous computing elements*** *that appears to its users as a single coherent system* - **Steen and Tanenbaum** in Distributed Systems: Principles and Paradigms, 3rd Edition, 2017

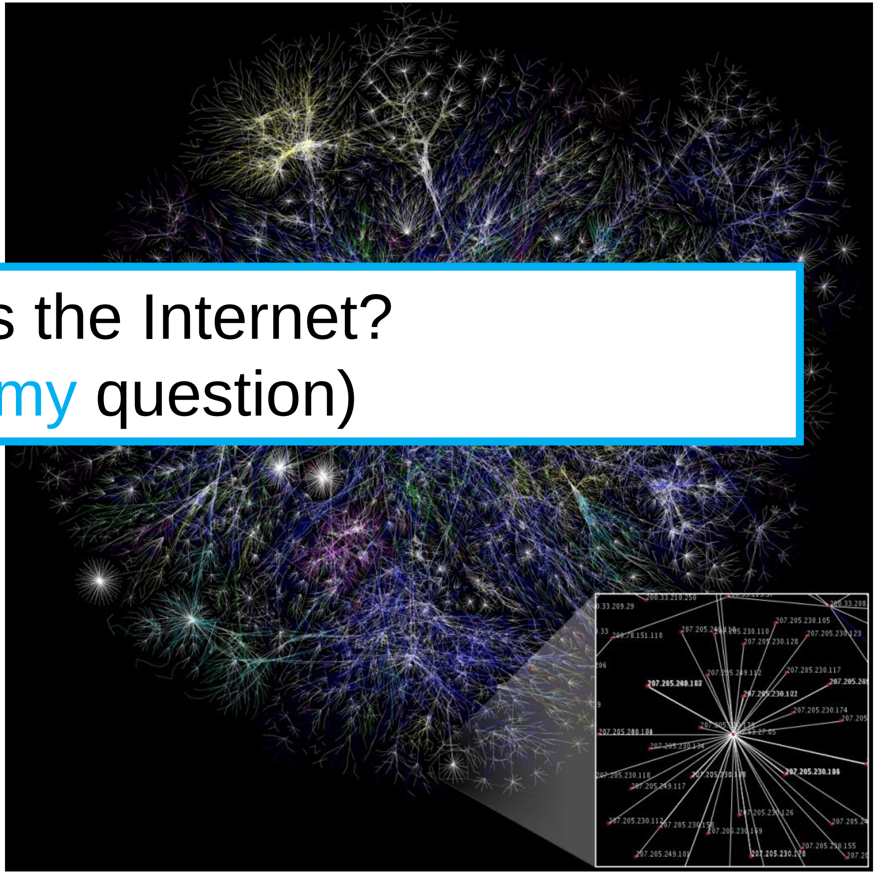
"an application that executes a collection of protocols to coordinate the actions of multiple processes on a network, such that all components cooperate together to perform a single or small set of related tasks."- Google University, Introduction to DS Design
http://www.hpcs.cs.tsukuba.ac.jp/~tatebe/lecture/h23/dsys/dsd-tutorial.html

VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

The Internet Is a Distributed System



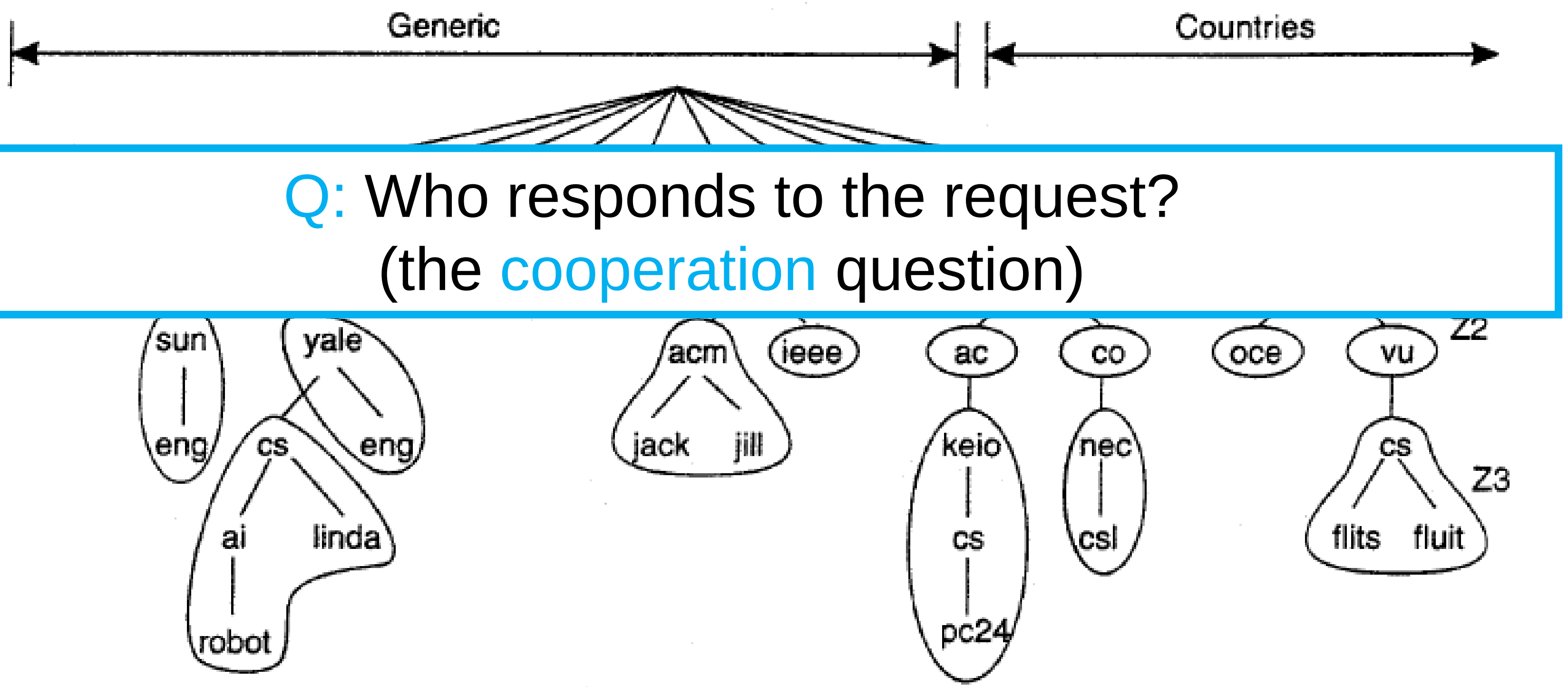


Q: Who owns the Internet?
(the autonomy question)

VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

The Domain Name System (DNS)



Q: Who responds to the request?
(the cooperation question)

The Google Data Centers

Q: How to compile the query answer?
(the communication question)

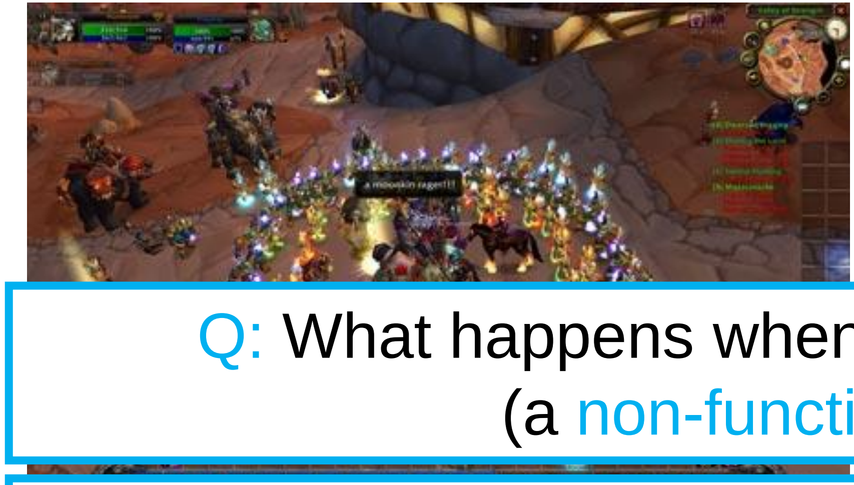
View our data centers in a larger map

Americas		Asia	Europe
Berkeley County, South Carolina	Mayes County, Oklahoma	Hong Kong	Hamina, Finland
Council Bluffs, Iowa	Lenoir, North Carolina	Singapore	St Ghislain, Belgium
Douglas County, Georgia	The Dalles, Oregon	Taiwan	Dublin, Ireland
Quilicura, Chile			

VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

The Online Gaming World

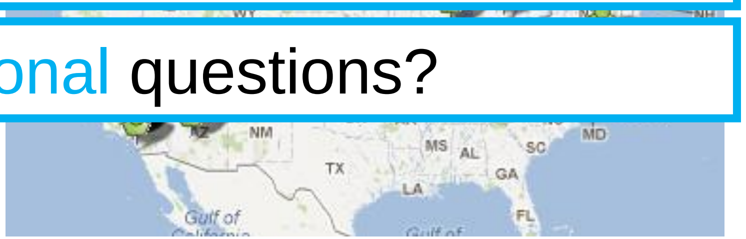


Q: What happens when the performance drops?
(a non-functional question)

Q: What happens when the servers are unavailable?
(another non-functional question)

Q: What other non-functional questions?

- 1.3PB storage
- 68 sysadmins (1/1,000 cores)



http://www.datacenterknowledge.com/archives/2009/11/25/wows-back-end-10-data-centers-75000-cores/

Agenda

1. What is a Distributed System?
2. **Distributed Systems, the Core Idea**
3. Distributed Systems, the Main Challenges
4. Relationship with Other Paradigms
5. Distributed Systems, a Design Example
6. Reality Check
7. Conclusion

The Core Idea through An Example

BitTorrent: A Distributed System

Q: Autonomy? Cooperation? Communication?

Q: Does this system scale? Why? How?

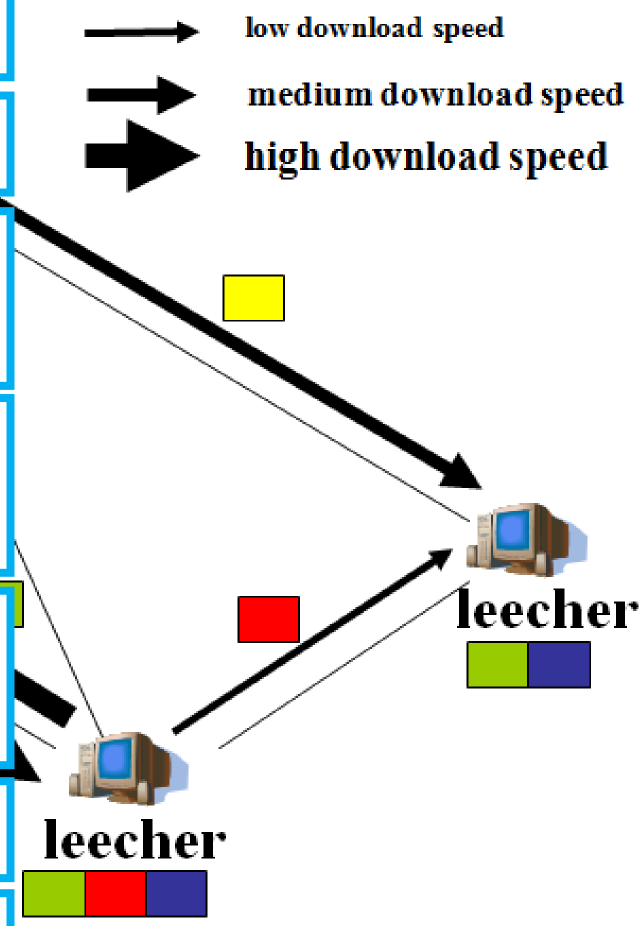
Q: What is the structure of this system?
What is the state of each node? How do they synchronize?

Q: How does the performance of this system change with the increase in the number of users?

Q: When is this system available?
What does it do to increase its reliability?

Q: Is this system efficient?

Q: Which parts of this system need consistency?
Achieved?



d.

Main Characteristics of Distributed Systems

1. Scalability
2. Predictable high performance
3. Reliability and availability
4. Efficiency (resource sharing)
5. Consistency of (distributed) state
6. Closeness-to-users
7. Transparency

So many other concerns:
Security,
Inter-operability, …

VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

Agenda

1. What is a Distributed System?
2. Distributed Systems, the Core Idea
3. **Distributed Systems, the Main Challenges**
4. Relationship with Other Paradigms
5. Distributed Systems, a Design Example
6. Reality Check
7. Conclusion

Main Challenges Raised by DS (1)

Q: Do DS have a regular, homogeneous

Q: Do DS have a commonly known state?

1. There is **not necessarily a regular structure**
   - **common protocols** for system components to cooperate
2. There is **no directly accessible common state** (this precludes shared-memory multiprocessors):
   - the system and applications need to **maintain a logical common state** by exchanging messages

VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

Main Challenges Raised by DS (2)

Q: Do DS have a common clock?

Q: Are DS deterministic (no randomness)?

Q: How do DS fail?

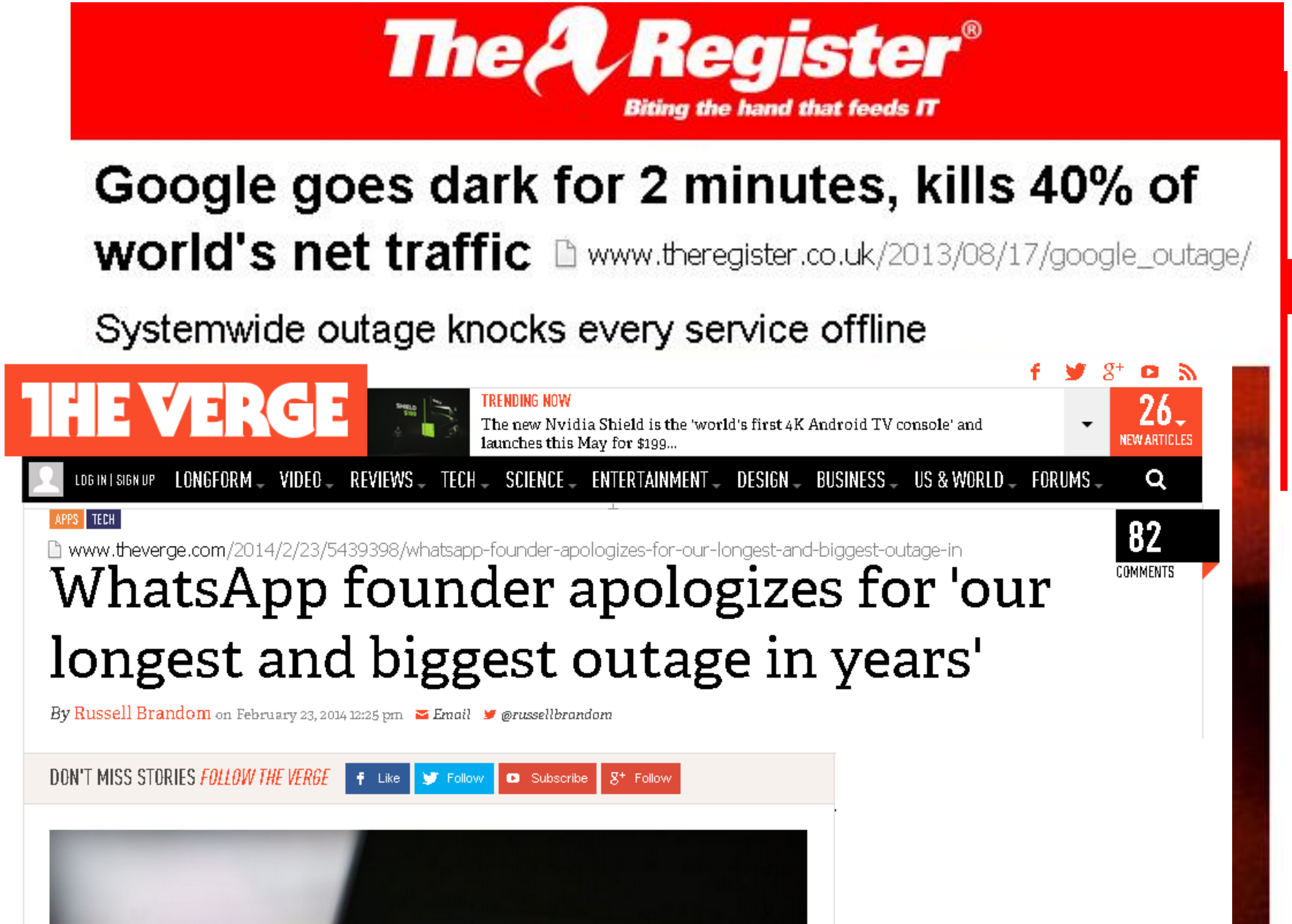
1. There is no common clock:
   - synchronization through the exchange of messages
2. There is non-determinism:
   - components make progress independently, often with users in the loop
   - operations can have side-effects on remote nodes
3. There are independent failure modes:
   - components may fail independently, and in many ways
   - failures are not observed immediately

VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

The Dependability* Challenge

* Availability, Reliability, etc.


The Register®
Biting the hand that feeds IT

Google goes dark for 2 minutes, kills 40% of world's net traffic
www.theregister.co.uk/2013/08/17/google_outage/

Systemwide outage knocks every service offline

THE VERGE
TRENDING NOW
The new Nvidia Shield is the 'world's first 4K Android TV console' and launches this May for $199...
26 NEW ARTICLES
LOG IN | SIGN UP LONGFORM VIDEO REVIEWS TECH SCIENCE ENTERTAINMENT DESIGN BUSINESS US & WORLD FORUMS
APPS TECH
82 COMMENTS
www.theverge.com/2014/2/23/5439398/whatsapp-founder-apologizes-for-our-longest-and-biggest-outage-in

WhatsApp founder apologizes for 'our longest and biggest outage in years'

By Russell Brandom on February 23, 2014 12:25 pm Email @russellbrandom

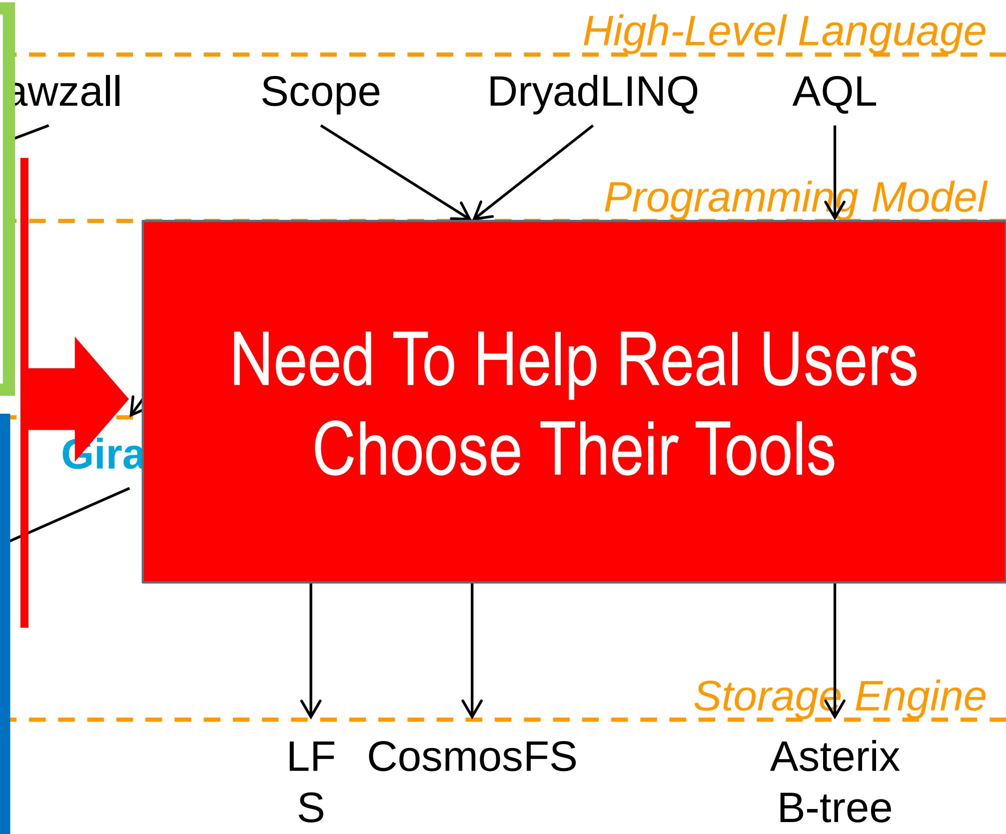
DON'T MISS STORIES FOLLOW THE VERGE Like Follow Subscribe Follow

Need Dependable Systems

The Ecosystem Navigation Challenge

System operator: how to prove capabilities? How to tune the tool? In which technology to invest? Which tech to DevOp in-house?

System customer: how to choose the right tool?
For batch, workflows, stream, transactions, etc.
(No one size fits all!)



Need To Help Real Users Choose Their Tools

Batch data processing ecosystem in 2011. A later example will cover the status in 2017.

The Scheduling Challenge

“30—70% scheduler decisions incorrect in datacenters”

Source: IEEE Computer’15

“current schedulers not efficient for many users, diverse services”

Source: Dutch industry, CCGRID’15

Need Smarter Schedulers

“new schedulers not used in datacenters, fear of failure”

Source: EuroPar’13,’14

Need to Select Schedulers

Agenda

1. What is a Distributed System?
2. Distributed Systems, the Core Idea
3. Distributed Systems, the Main Challenges
4. **Relationship with Other Paradigms**
5. Distributed Systems, a Design Example
6. Reality Check
7. Conclusion

Distributed vs Parallel Computing

Distributed	Parallel Computing
• Multiple tasks, one job or multiple jobs	• Multiple tasks, one job
• Throughput or Speed-up • Horizontal scaling	• Speed-up • Vertical scaling
• Infrequent communication • Synchronized execution	• Frequent communication • Simultaneous execution
• Heterogeneous hardware	• Homogenous hardware
• Multiple owners with mutual interests	• Single owner

Q: High Performance Computing?

Q: Cluster of GPUs?

Q: GPU processing?

Q: Cluster computing?

Distributed Variants

- Most grid computing
- Most cloud computing
- Peer-to-Peer computing
- Most Big Data processing (MapReduce/Hadoop2, Pregel/Giraph, Spark, etc.)
- Cluster computing
- Some High-Performance Computing

Agenda

1. What is a Distributed System?
2. Distributed Systems, the Core Idea
3. Distributed Systems, the Main Challenges
4. Relationship with Other Paradigms
5. **Distributed Systems, a Design Example**
6. Reality Check
7. Conclusion

Distributed Systems, a Design Example
2Fast: Collaborative Downloading

“In two years time we will all have petabytes on our key chains and will not need BitTorrent at all”
(anonymous, for the sake of this course, 2005)

P. Garbacki, A. Iosup, D.H.J. Epema, and M. van Steen, "2Fast: Collaborative Downloads in P2P Networks," *6-th IEEE International Conference on Peer-to-Peer Computing*, 2006 (**best-paper award**).

VRIJE UNIVERSITEIT AMSTERDAM

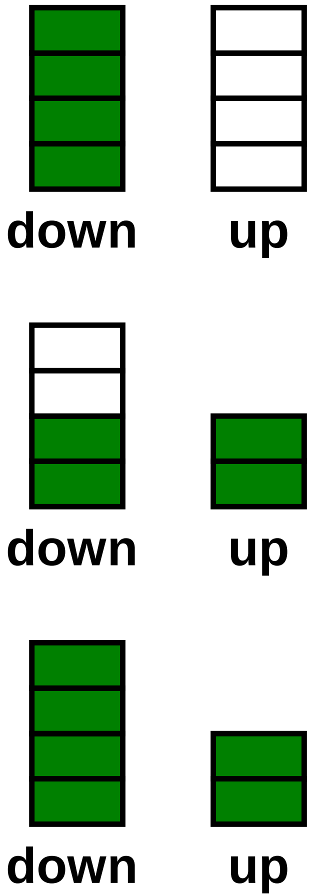
TU Delft

Peer-to-peer data transfer protocols

- Gnutella, Kazaa
  - no incentives for bandwidth sharing
  - free-riders sensitive
  - poor utilization of upload bandwidth

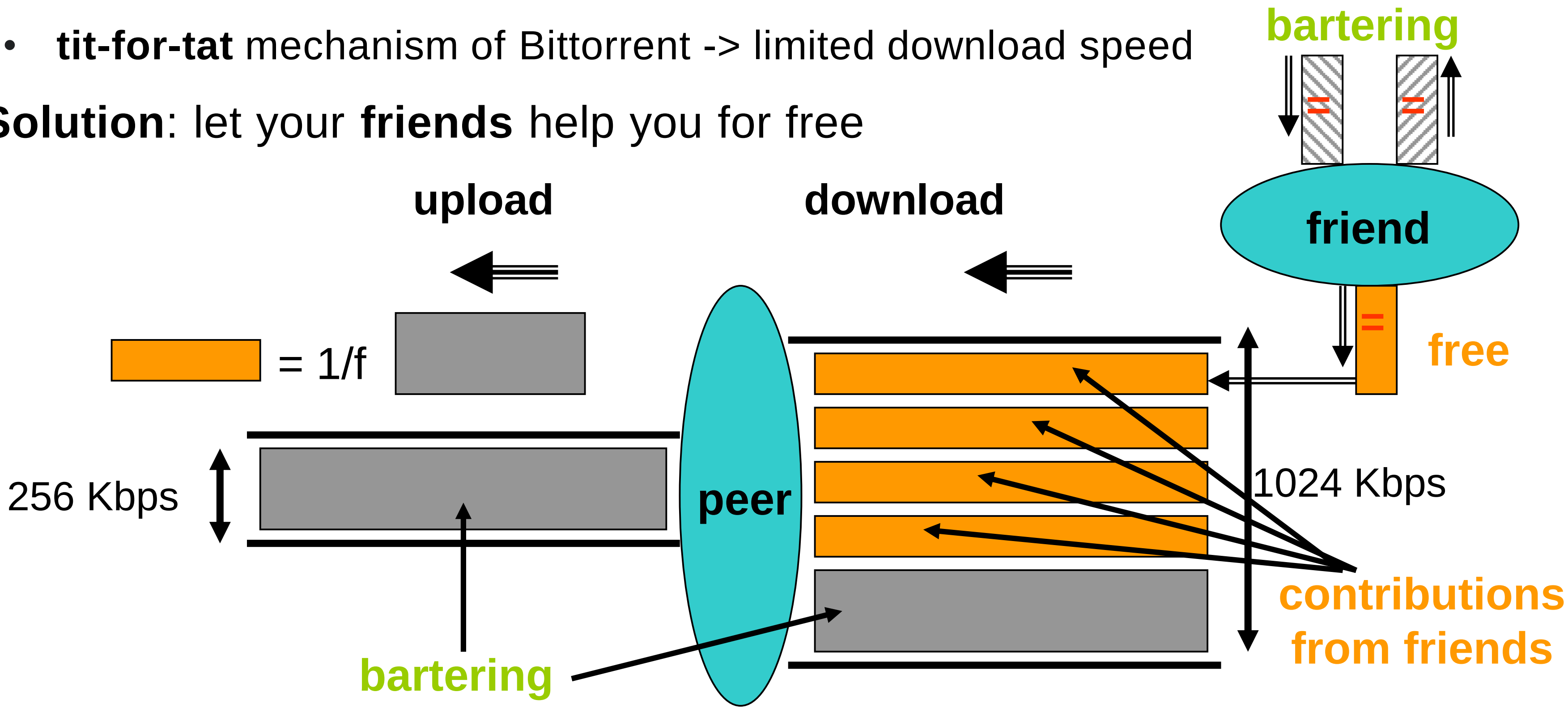
- BitTorrent (BT), Slurpie
  - tit-for-tat enforces fairness
  - temporal fairness cannot handle asymmetric links
  - poor utilization of download bandwidth

- **2Fast: BT+collaborative downloads**
  - no tit-for-tat within a single session
  - cross-session bandwidth sharing
  - full utilization of upload AND download links



Cooperative downloads: basic idea

- **Problem**:
  - most users have **asymmetric** upload/download links
  - **tit-for-tat** mechanism of Bittorrent -> limited download speed
- **Solution**: let your **friends** help you for free



Two protocol extensions

- **Redundant chunks download**
  - **problem**: discrimination of helpers; more restrictive chunk selection + fewer chunks to offer, so limited bartering possibilities
  - **solution**: the same chunk may be downloaded by different helpers
- **Sharing of swarm information**
  - **problem**: slow start; finding suitable bartering partners takes time
  - **solution**: collaborating peers exchange information on other peers in the swarm

Download speed-up: analytical model

- Every helper **equally splits its upload capacity** between bartering and helping the collector
- So **every additional helper** increases the download speedup of the collector by 0.5, up to a point
- The **maximum number of useful helpers** (and so the maximum speedup) can easily be computed
- **Download bandwidth** of the collector with **h helpers**:
  - N, S: the numbers of **leechers** and **seeders** in the system
  - c, μ: the download/upload capacity of all peers (**homogeneous** model)

free from seeders $\boxed{\frac{S}{N}\mu} + \boxed{\mu} + \boxed{\frac{1}{2}\sum_{i=1}^{h}(\frac{S}{N}+1)\mu}$ from helpers

bartering

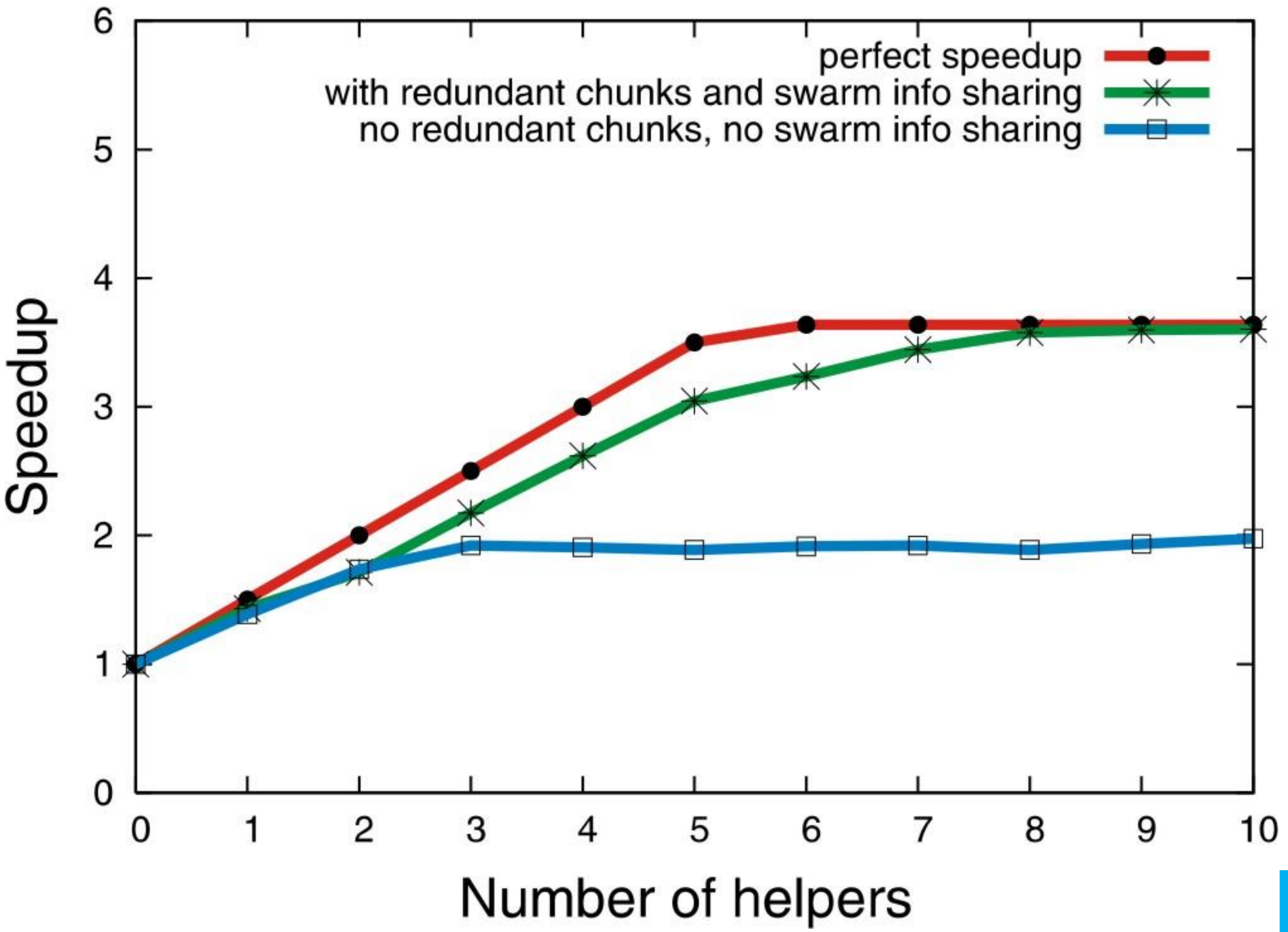
Experimental setup: DAS + real-world

- Experiments performed in a real environment – collaborating peers connect to existing BitTorrent swarms
- Collaborating peers connected through ADSL links: 256kbps up / 1024kbps down
- Downloaded file size: 700 MB
- Swarm size: 100 leechers, 10 seeders

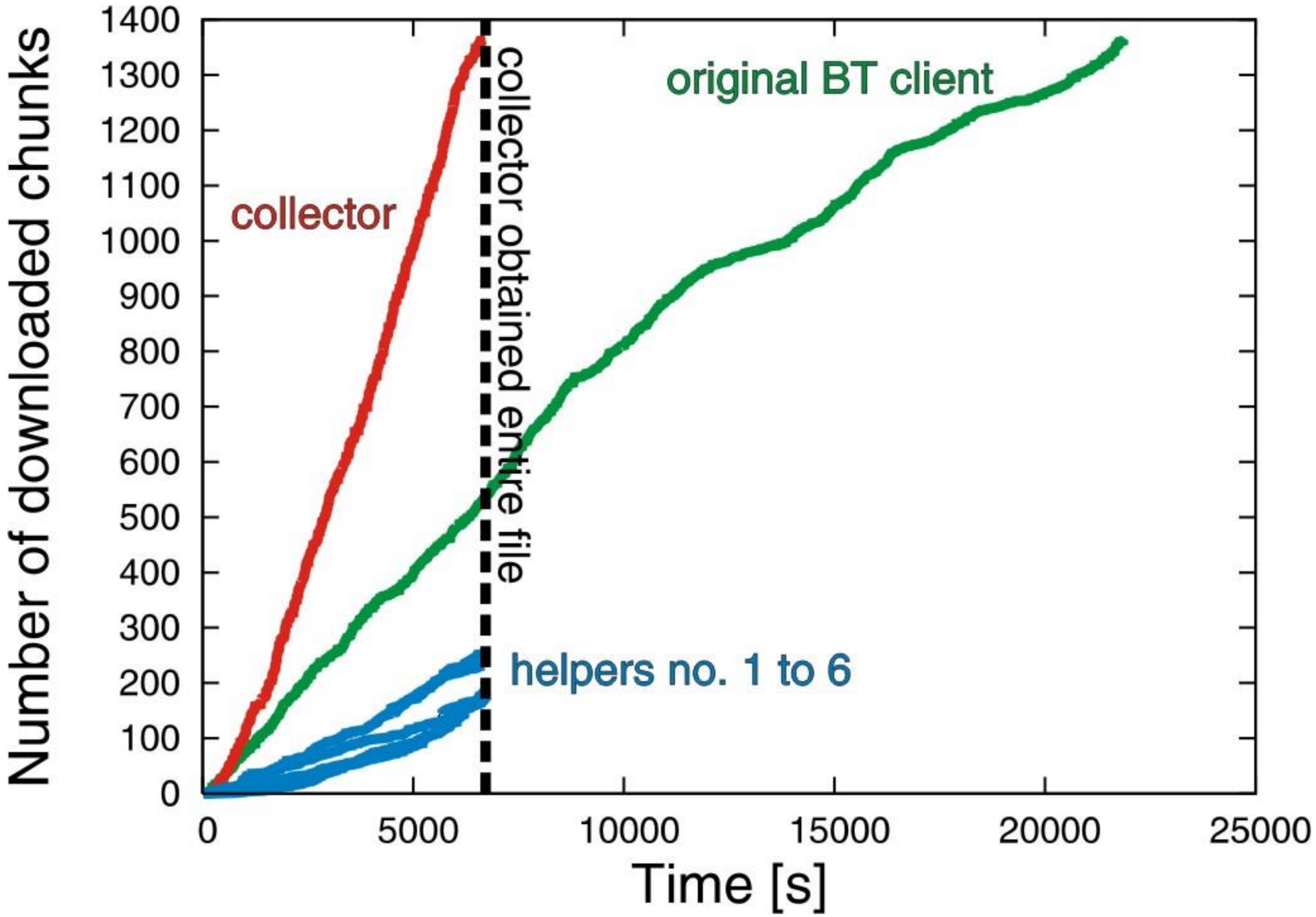
VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

Speedup vs number of helpers



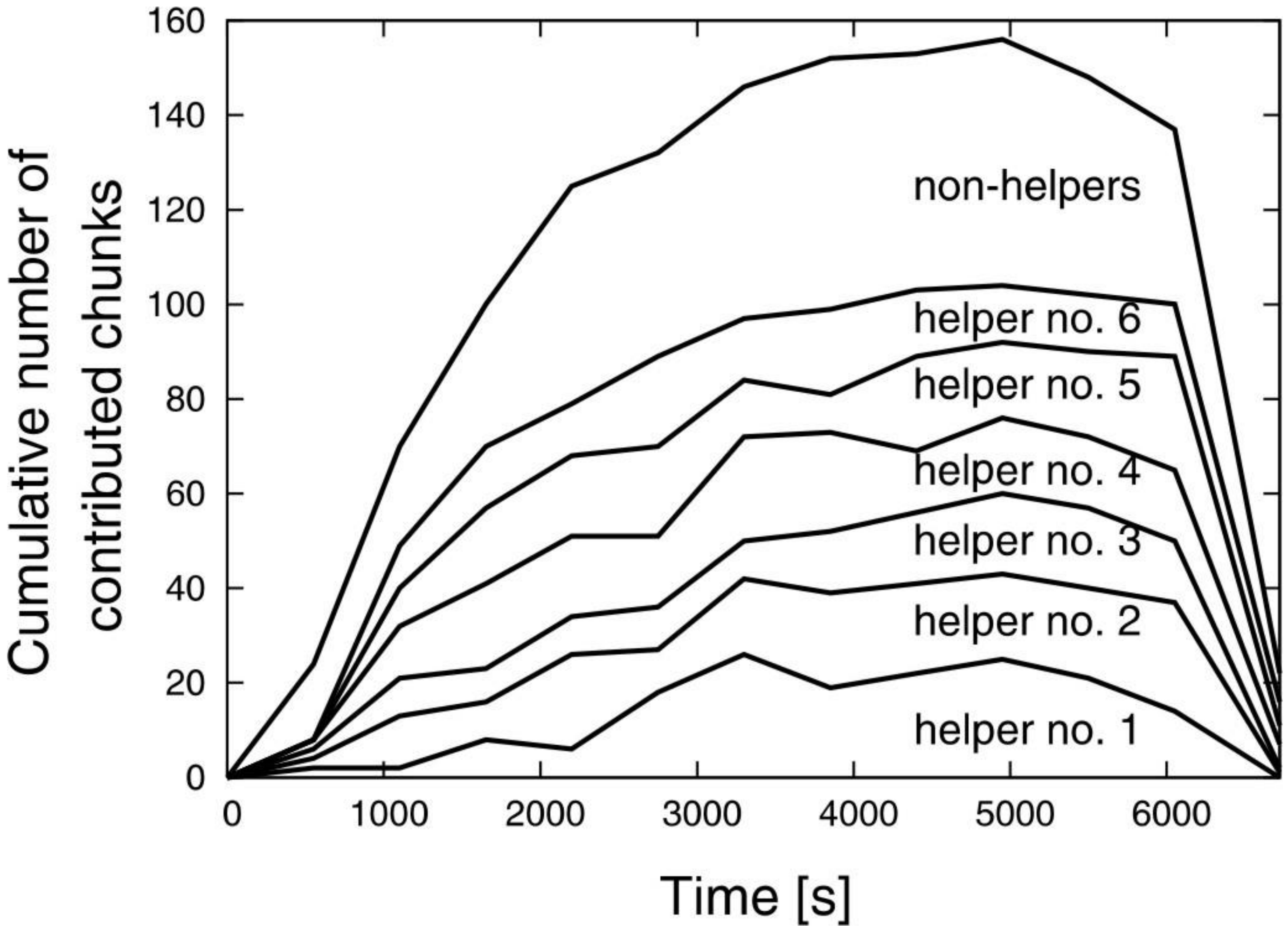
Download progress



VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

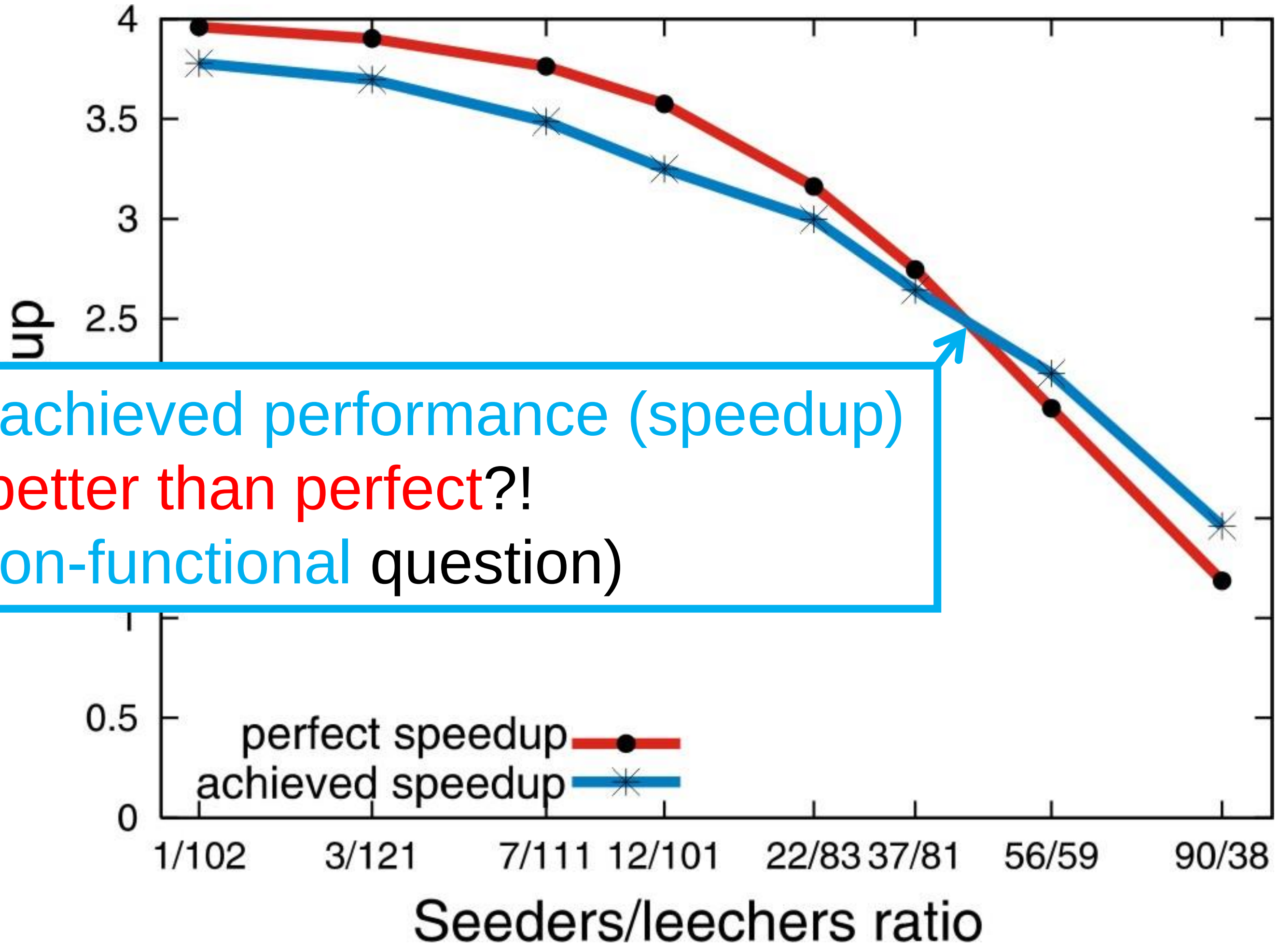
Helper contributions over time



VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

Speedup vs. seeders/leechers ratio



Q: Why is the achieved performance (speedup) better than perfect?!
(a non-functional question)

the more seeders, the more bandwidth for free, and so the less benefit from helpers

VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

Agenda

1. What is a Distributed System?
2. Distributed Systems, the Core Idea
3. Distributed Systems, the Main Challenges
4. Relationship with Other Paradigms
5. Distributed Systems, a Design Example
6. **Reality Check**
7. Conclusion

VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

Tools and Companies

All logos belong to companies, as copyrighted or trade-marked branding elements.

Active Research Field

- **Research** in the Massivizing Computer Systems group:
  www.atlarge.science
- Symposium on High Performance Distributed Computing (HPDC)
  - www.informatik.uni-trier.de/~ley/db/conf/hpdc/index.html
- Symposium on Networked Systems Design and Implementation (NSDI)
  - www.informatik.uni-trier.de/~ley/db/conf/nsdi/
- Symposium on Cluster Computing and the Grid (CCGRID)
  - www.informatik.uni-trier.de/~Ley/db/conf/ccgrid/index.html
- IEEE Transactions on Parallel and Distributed Systems
  - www.informatik.uni-trier.de/~ley/db/journals/tpds/

VU VRIJE UNIVERSITEIT AMSTERDAM

TU Delft

Fundamental Research in Massivizing Comp. Sys.

Scheduling
Bags-Of-Tasks
Workflows
Portfolio

Dependability
Failure Analysis*
Space-/Time-Correlation
Availability-On-Demand

New World+
Workload Modeling
Business-Critical
Online Gaming

Ecosystem Navigator+
Performance Variability
Grid*, Cloud, Big Data
Benchmarking*
Longitudinal Studies

Scalability/Elasticity+
Delegated Matchmaking*
BTWorld*, POGGI*, AoS
Auto-Scalers
Heterogeneous Systems

Socially Aware+
Collaborative Downloads*
Groups in Online Gaming
Toxicity Detection*
Interaction Graphs

Education
Social Gamification*

Software Artifacts
Graphalytics, OpenDC

Data Artifacts
Distributed Systems Memex*

Fundamental Problems/Research Lines
My Contribution So Far
Personal grants

+ Please ask for a definition
* Award-level

Agenda

1. What is a Distributed System?
2. Distributed Systems, the Core Idea
3. Distributed Systems, the Main Challenges
4. Relationship with Other Paradigms
5. Distributed Systems, a Design Example
6. Reality Check
7. **Conclusion**

~~Conclusion~~ Take-Home Message

- **Distributed Systems = autonomy + cooperation + communication**

- **Core idea = autonomous nodes using communication to cooperate**
    - Scalability, Resource sharing, Reliability and availability, Predictable high performance, Consistency, Close-to-Users, …

- **Reality Check: we are all users**
  Google, Facebook, Twitter, netflix, …

http://www.flickr.com/photos/dimitrisotiropoulos/4204766418/

VU VRIJE UNIVERSITEIT AMSTERDAM

TUDelft

Entry Quiz

(closes after class)

- You choose if you want to do this quiz
  - Not mandatory
  - 500p at stake

The images used in this lecture courtesy of the Computer History Museum, Mountain View, California, USA, http://www.computerhistory.org/ ; the German Museum of Technology (Deutsches Technikmuseum Berlin, Germany, http://www.sdtb.de/Englisch.55.0.html ; the Science Museum, London, UK, http://www.sciencemuseum.org.uk/; and many anonymous contributors via Google Images. Many thanks!