SmartSockets:

Solving the Connectivity Problems in Grid Computing

Jason Maassen
jason@cs.vu.nl

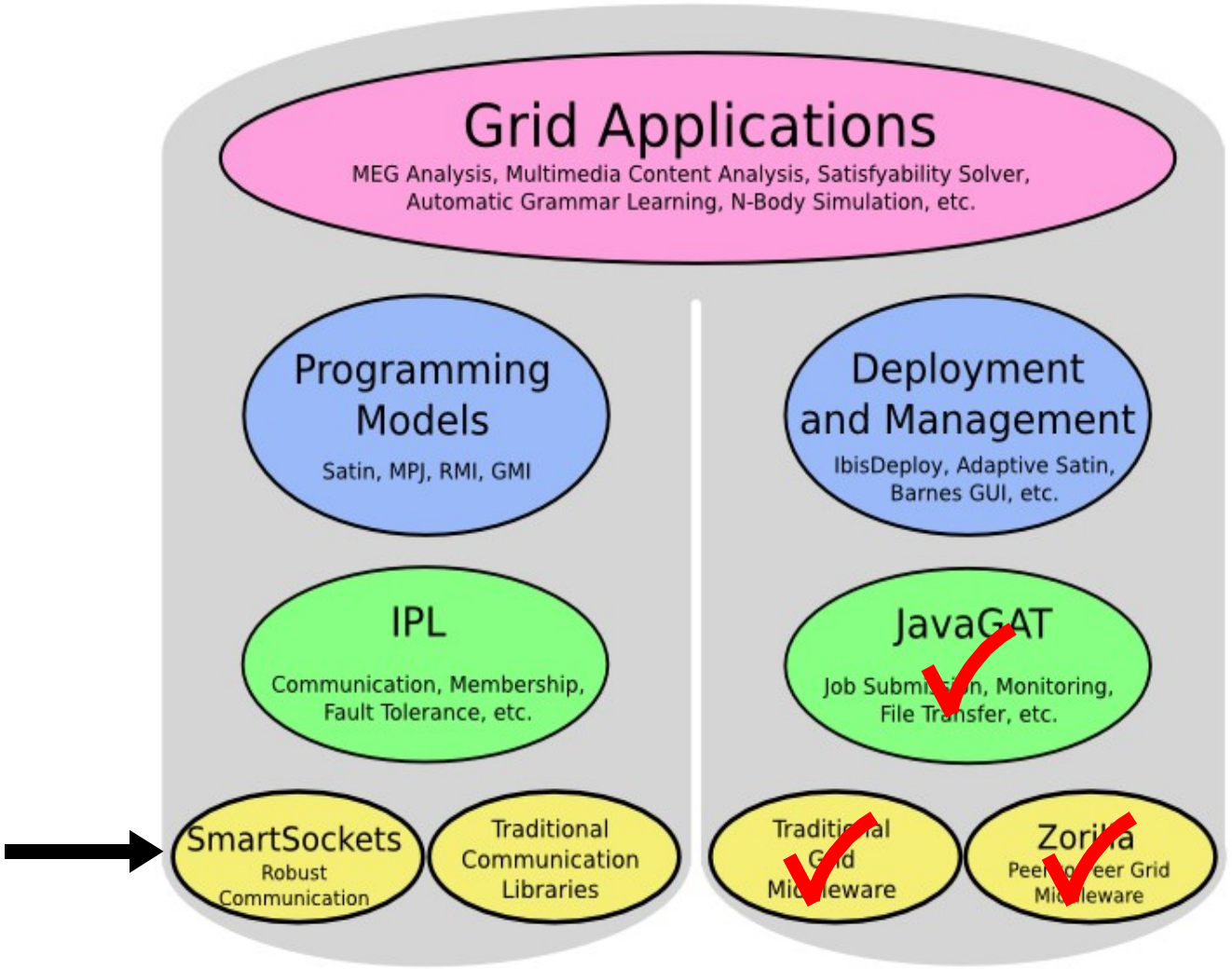
vrije Universiteit

Until now...

- We looked at how Ibis makes Grid middleware user friendly:
  - **JavaGAT** provides an easy-to-use API for the various flavours of Grid middleware
  - **Zorilla** provides a configuration free alternative to existing Grid middleware
- Next step:
  - **Communication**

vrije Universiteit

Overview



ibis

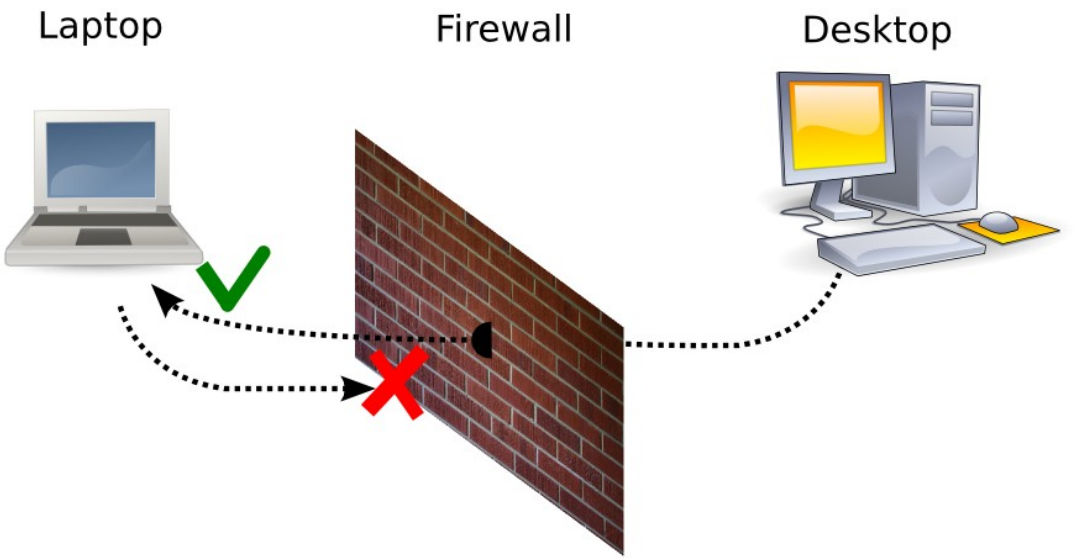
vrije Universiteit

Communication is Difficult!

- Many sites have connectivity issues
  - Firewalls
  - Network Address Translation (NAT)
  - Non-routed networks
  - Multi homing
  - Mis-configured machines
  - …
- This makes it hard to use a combination of sites

Problem 1: Firewalls

- Blocks 'inapropriate' traffic
  - Usually only blocks incoming traffic
  - Some also block outgoing traffic

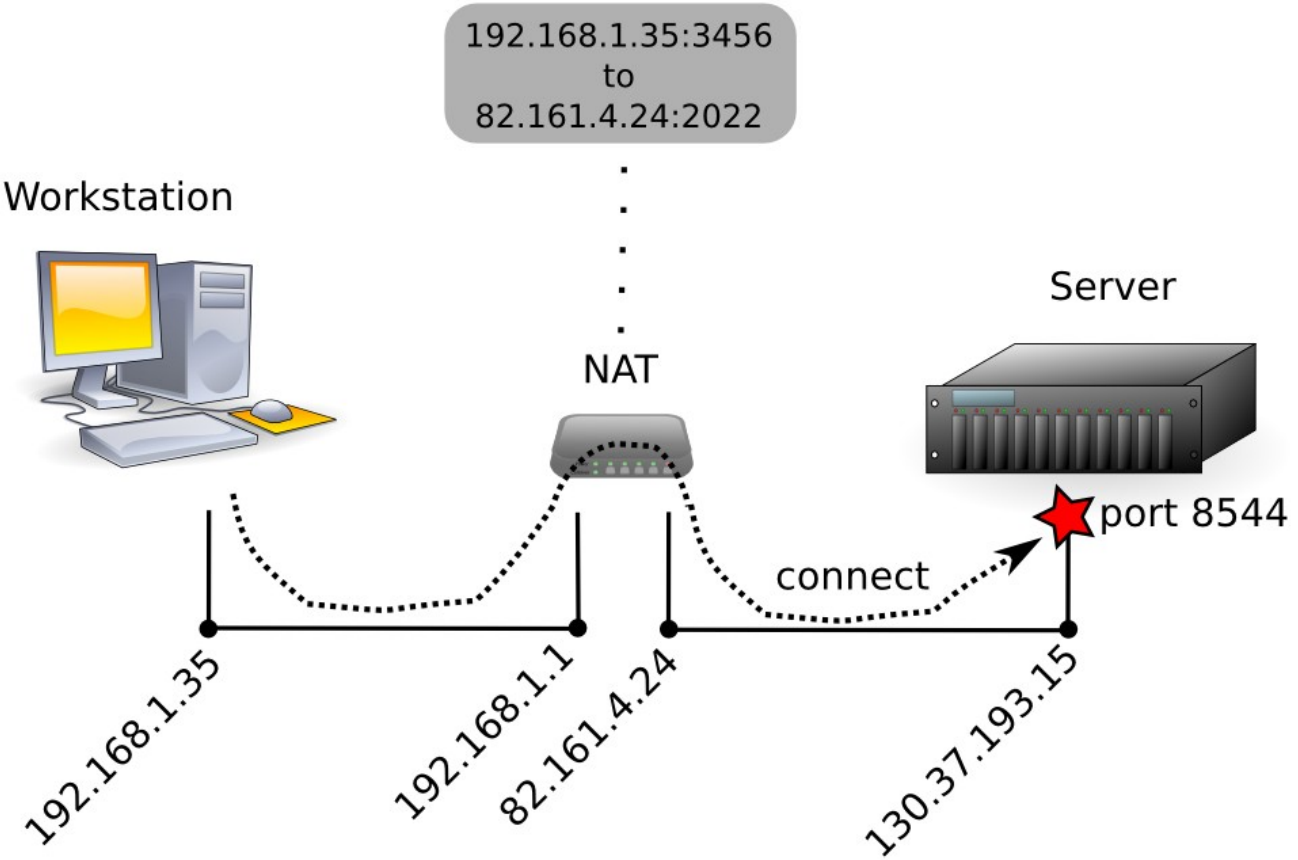


ibis

vrije Universiteit

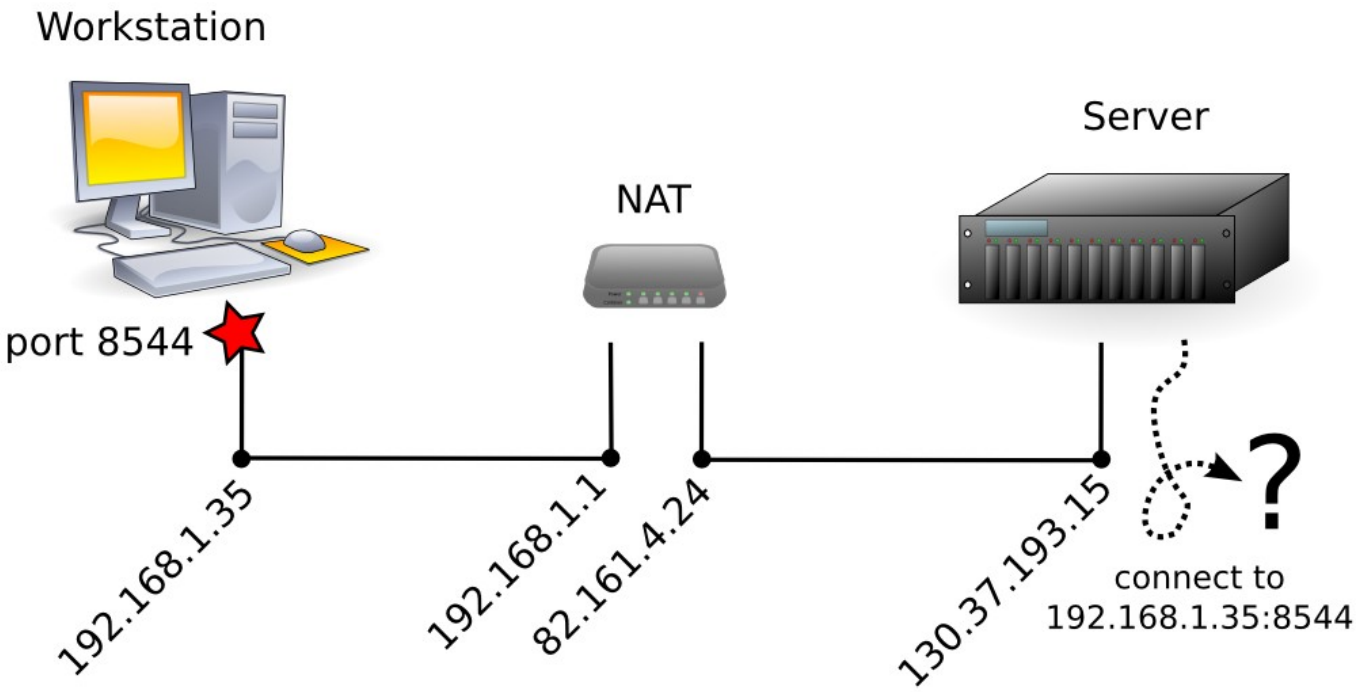
Problem 2: Network Address Translation

- Maps IP from one range into another
  - Works fine for outgoing connections



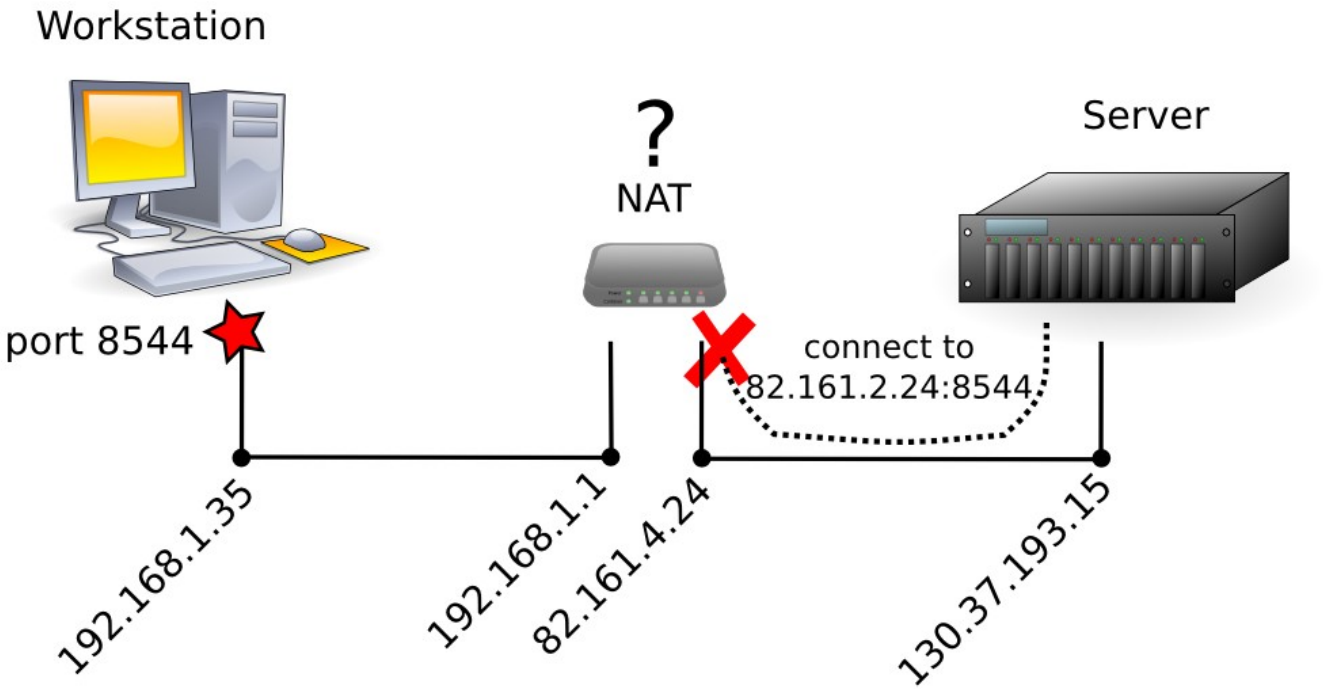
Problem 2: Network Address Translation

- Problem: incoming connections
  - Target address is invalid on internet



Problem 2: Network Address Translation

- Problem: incoming connections
  - Target address is invalid on internet
  - NAT device does not know where to forward connection

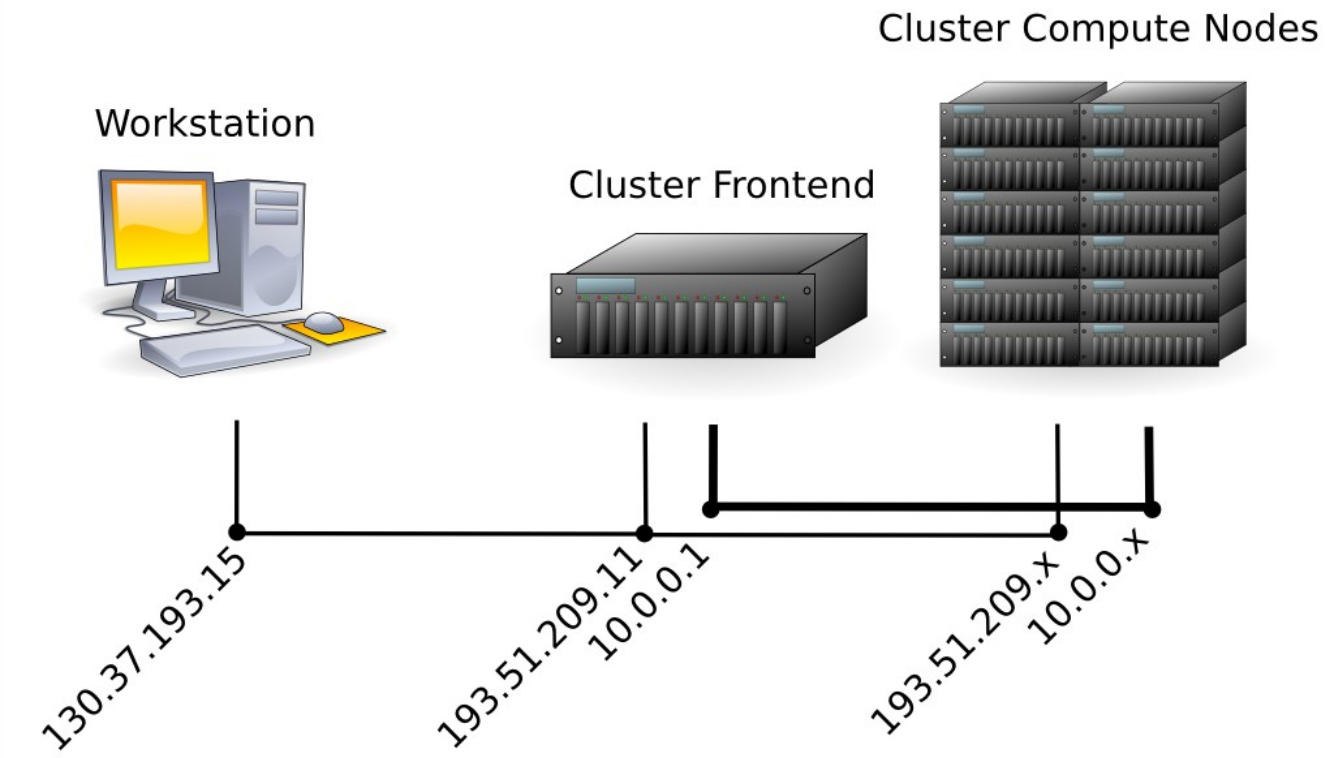


ibis

vrije Universiteit

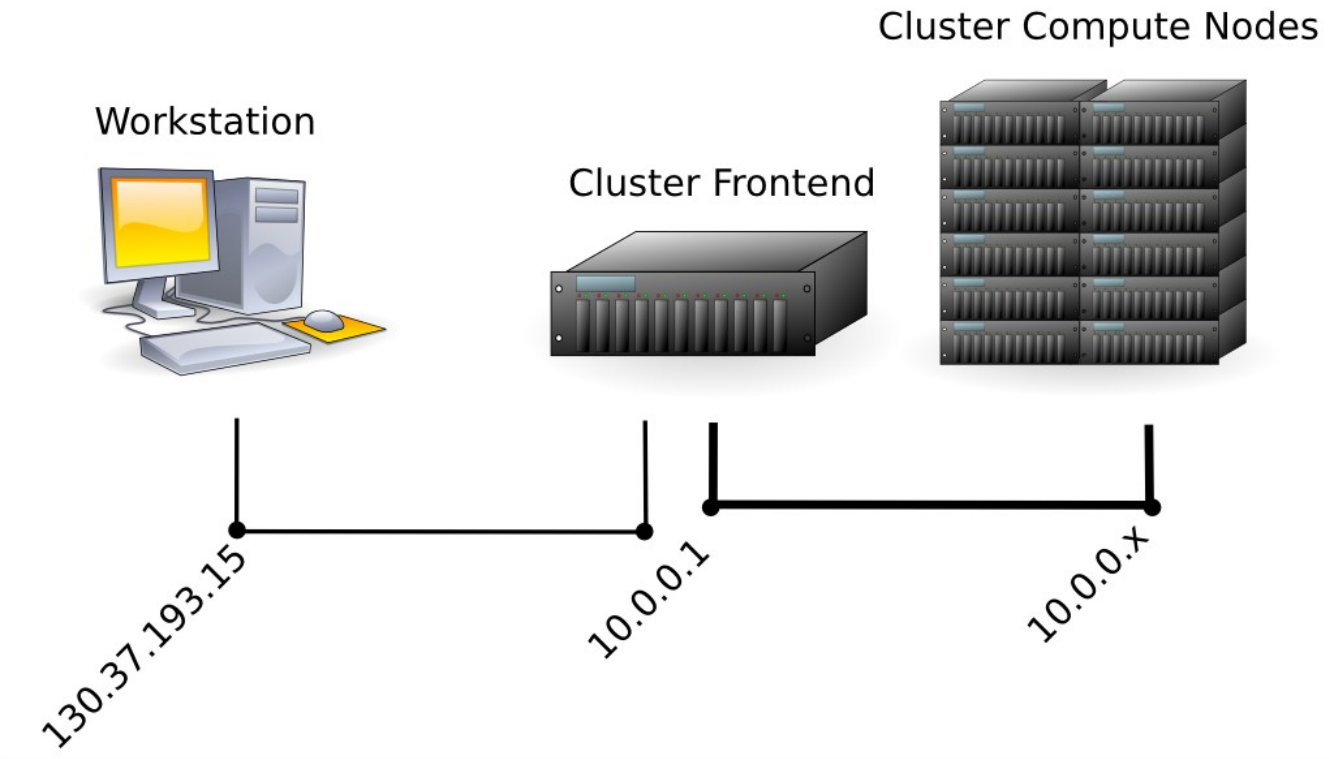
Problem 3: Multi Homing

- Some sites have multiple networks
  - The target address depends on the source of the connection



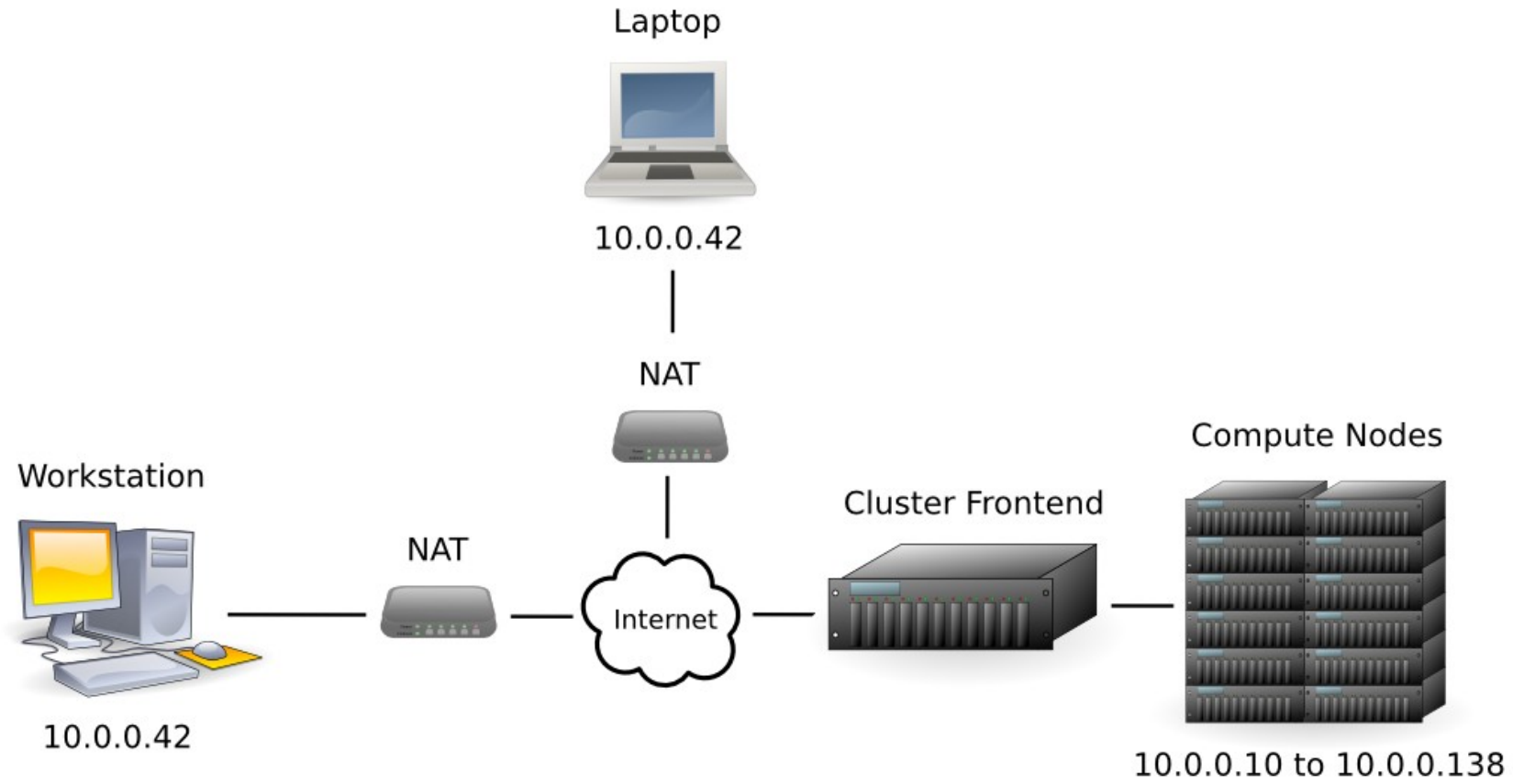
Problem 4: Non-routed Networks

- Some sites do not route between the local network and internet
  - Only the frontend is reachable



Problem 5: Machine Identification

- NAT/non-routed networks lead to machine identification problems



Current Solutions

- There are many ad-hoc solutions:
  - Open port ranges in the firewall
  - Use port forwarding in NAT
  - Explicitly specifying IP addresses
  - SSH tunneling
  - ...
- These solutions work (sort of), but...

Current Solutions (Cont'd)

- It all up to the user!
  - What is wrong ?
  - How it can be solved ?
- A lot of work
  - Need help from system administration
  - May need changes to application
  - Needs to be re-done whenever the testbed changes!

SmartSockets

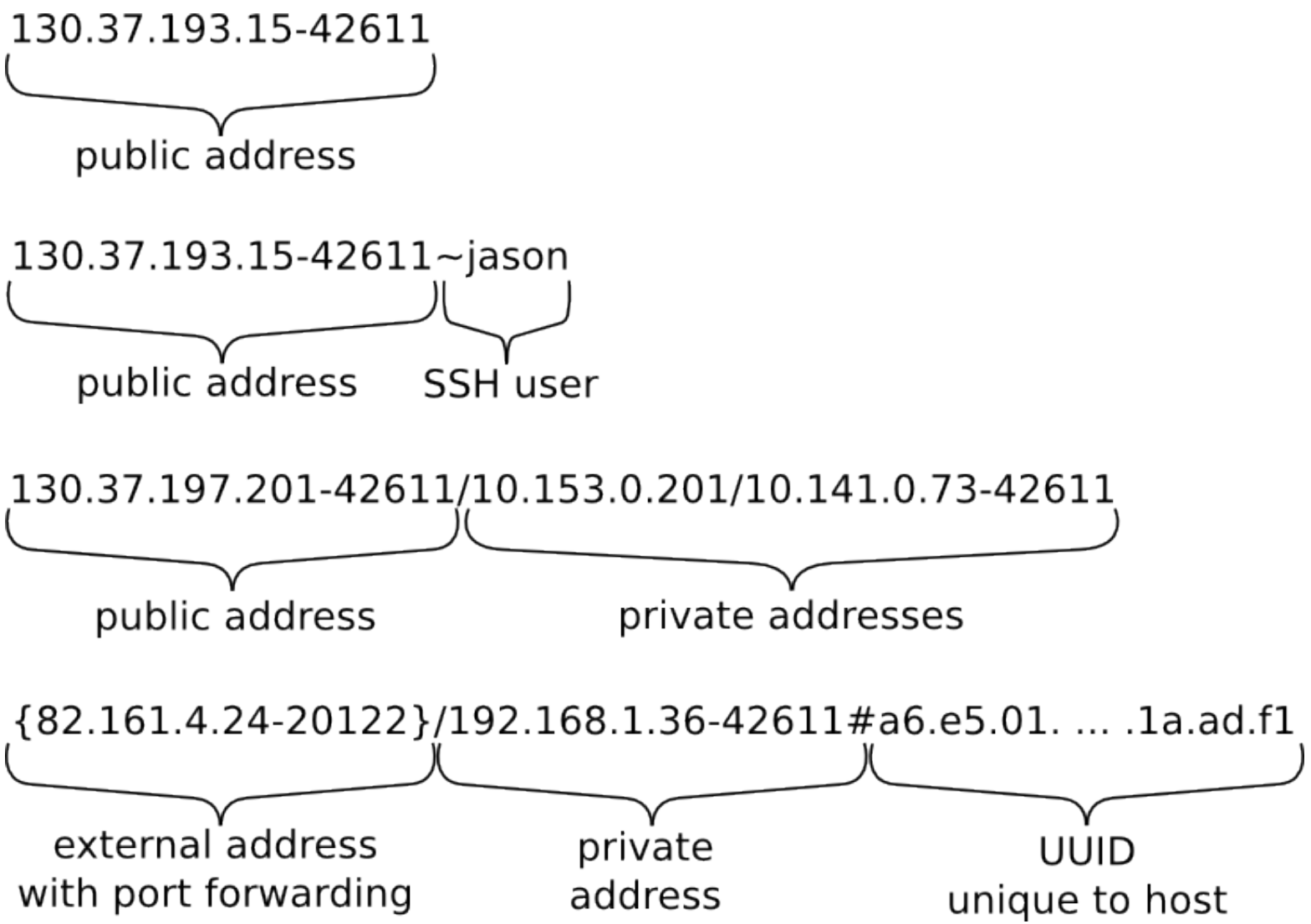
- The SmartSockets library
  - Offers a socket-like interface
    - Addressing is different
  - Detects connectivity problems
  - Tries to solve them automatically
    - With as little help from the user as possible
  - Integrates existing and several new solutions into one library
- User friendly connection setup!

vrije Universiteit

Smart Addressing

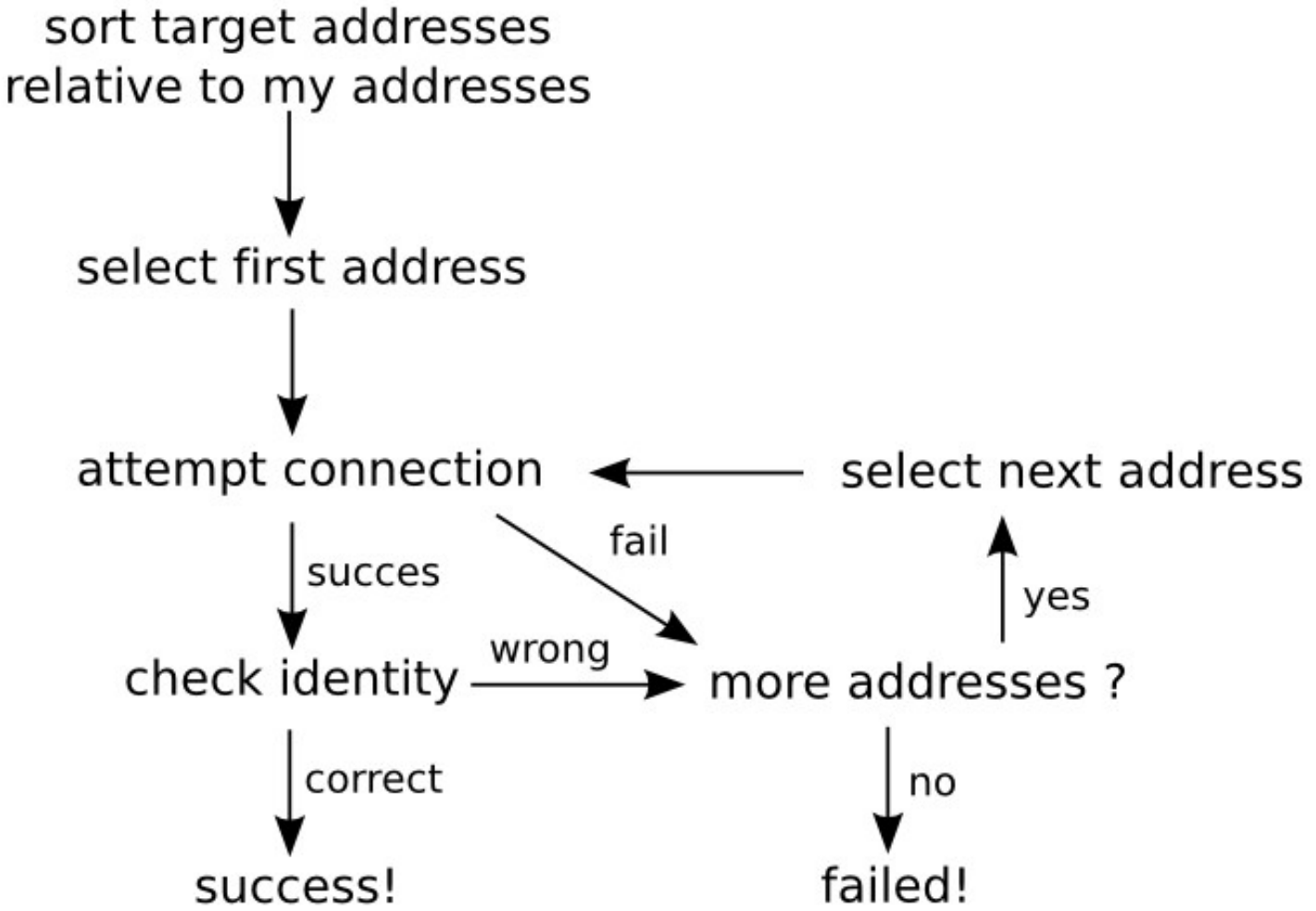
- SmartSockets extends addressing
  - Not just a single IP:port combination
- Add all machine addresses
- Optionally add extra information
  - External address + port (for NAT)
  - UUID (if entire address is private)
  - SSH contact information

Addressing Examples



vrije Universiteit

Creating a Connection



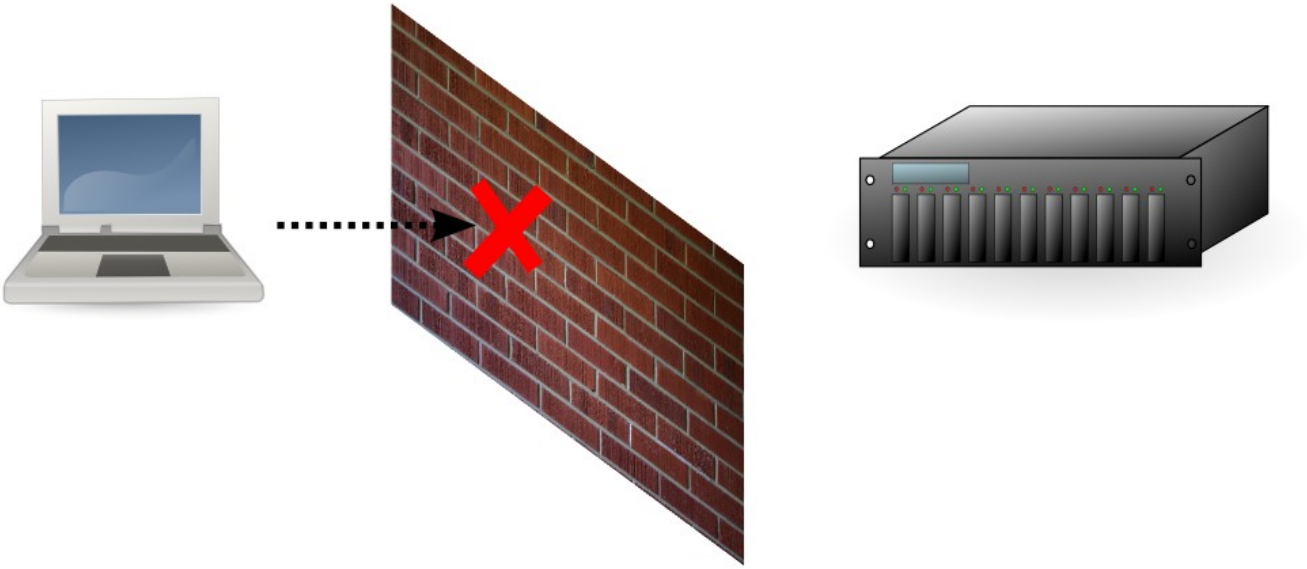
ibis

vrije Universiteit

First Results

- This solves multi homing and machine identification problems (3 & 5)
- Assumes anyone can create a connection
  - This will not help when the target is behind a NAT / Firewall
- To solve this we need cooperation between machines....

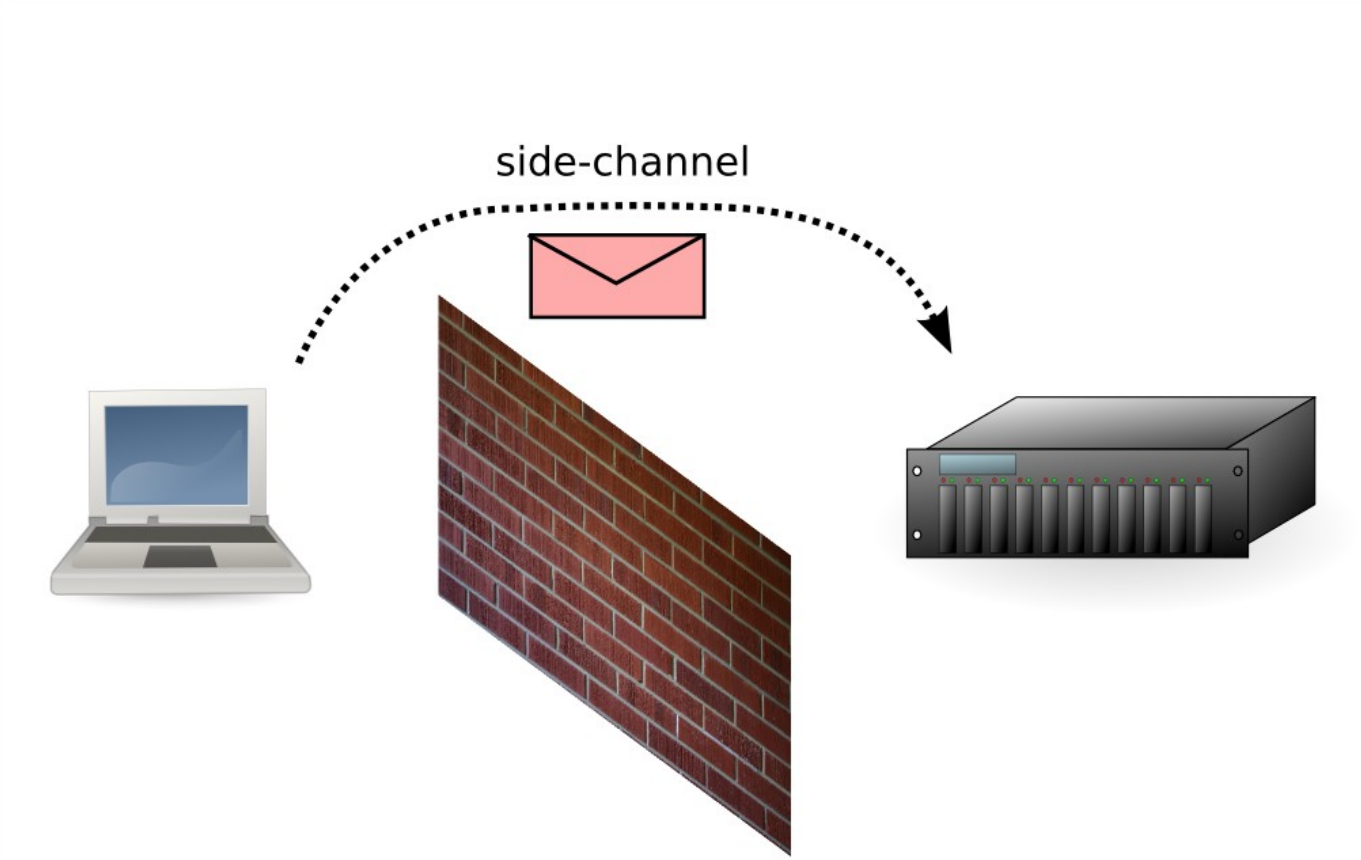
Cooperation



ibis

vrije Universiteit

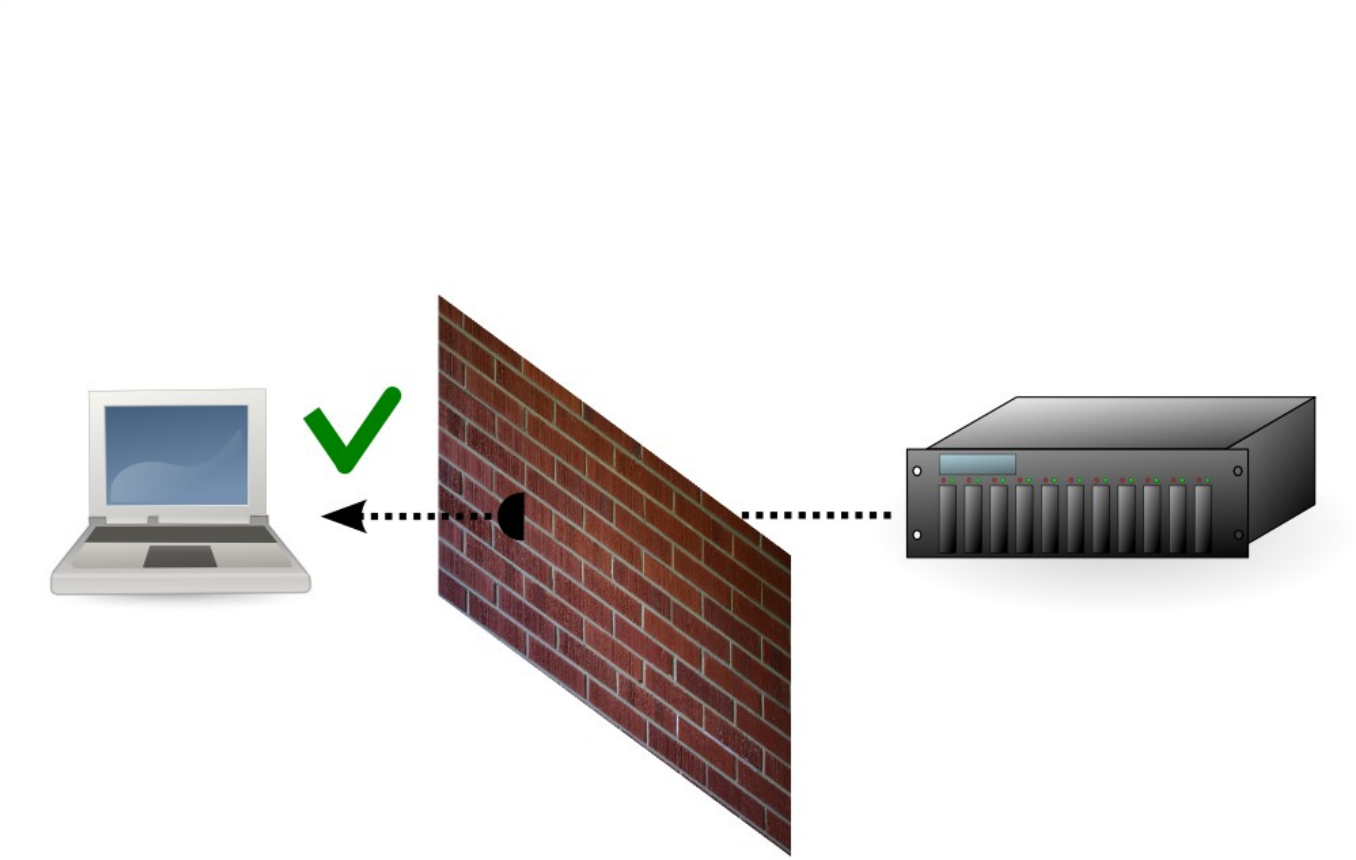
Cooperation



ibis

vrije Universiteit

Cooperation



ibis

vrije Universiteit

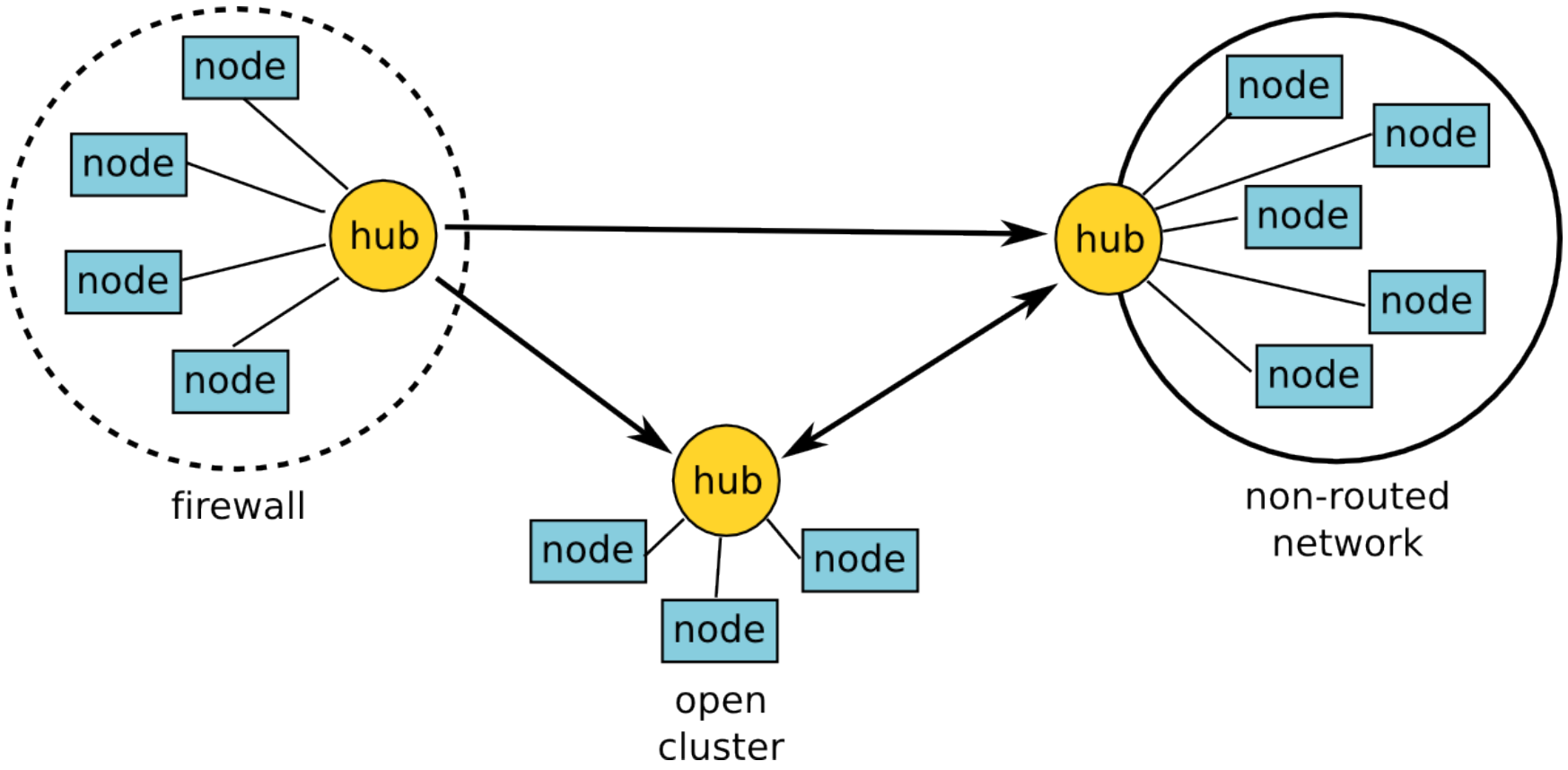
Side Channel

- SmartSockets uses network of *hubs* to implement a *side channel*
  - Support processes for the application
  - Started in advance
- Hubs are run on machines with 'more connectivity'
  - Such as cluster frontends, 'open' machines, etc.

Hubs

- Hub connect to each other
  - Need to set up spanning tree (or better)
    - Use direct connections and SSH tunnels
  - Gossip information and client messages
    - Automatically discover new hubs and routes
- Clients connect to a 'local' hub
  - When needed, use network of hubs as side channel for connection setup

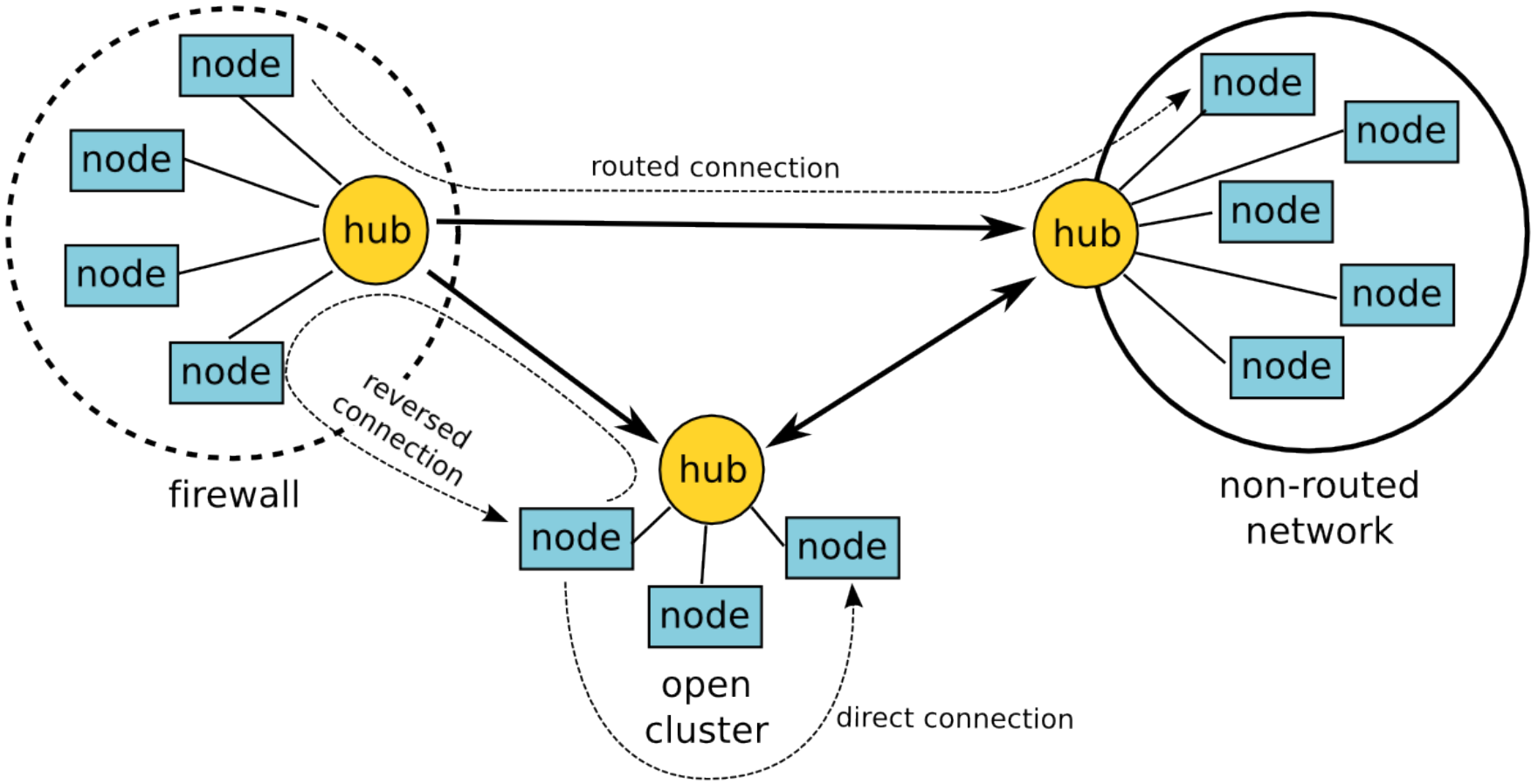
Example



ibis

vrije Universiteit

Example Cont'd



ibis

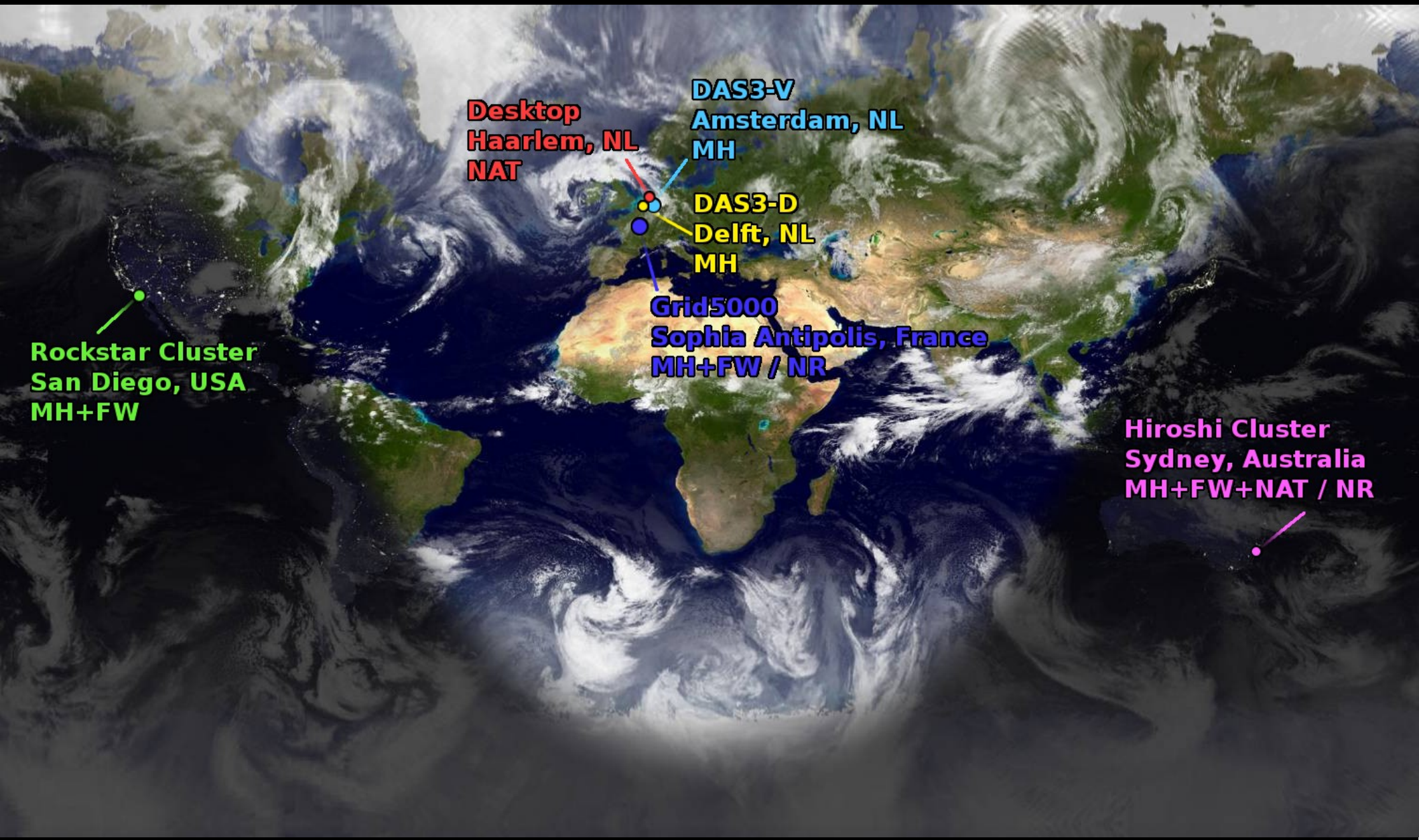
vrije Universiteit

Reverse

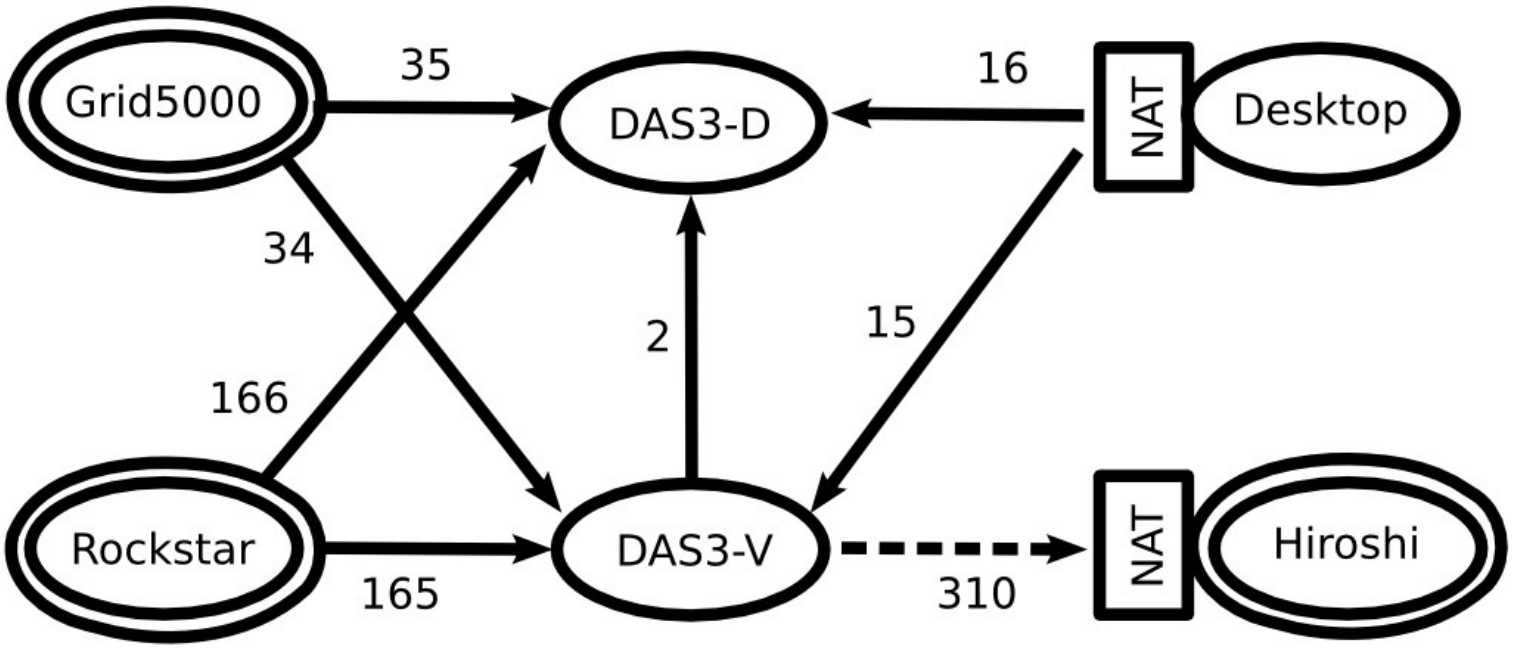
- Reverse direction of connection setup
  - Send message to target using hub
  - Wait for incoming (direct) connection
- Results in **direct connection**
  - Only connection setup is different
- Solves Firewall/NAT problems (1 & 2)
  - Assumes one side can create connection
  - Does not work with multiple firewalls or NATs, or non-routed networks

Routed

- Create *virtual connection* using hubs
  - Forward all data over side channel
- Results in **indirect connection**
  - Usually lower performance
  - Still looks like a socket!
- Solves non-routed problem (4)
  - Also solves the multi-firewall/NAT case


Desktop
Haarlem, NL
NAT
DAS3-V
Amsterdam, NL
MH
DAS3-D
Delft, NL
MH
Grid5000
Sophia Antipolis, France
MH+FW / NR
Rockstar Cluster
San Diego, USA
MH+FW
Hiroshi Cluster
Sydney, Australia
MH+FW+NAT / NR

Hub Network



ibis

vrije Universiteit

Evaluation

Table 3: Connection setup time of SmartSockets (time in milliseconds).

	Source					
Target	DAS3-V	DAS3-D	Rockstar	Grid5000	Hiroshi	Desktop
DAS3-V		4.9^{d} (2.4)	332^{d} (166)	68^{v}	595^{v}	33^{d} (17)
DAS3-D	4.9^{d} (2.4)		335^{d} (167)	70^{v}	595^{v}	33^{d} (18)
Rockstar	500^{r}	503^{r}		206^{v}	718^{v}	182^{v}
Grid5000	35^{v}	38^{v}	206^{v}		593^{v}	54^{v}
Hiroshi	630^{v}	603^{v}	750^{v}	670^{v}		640^{v}
Desktop	49^{r}	52^{r}	183^{v}	84^{v}	606^{v}	

Annotations indicate connection style: d for direct, r for reverse, s for splicing, and v for routed. When applicable, the connection setup time of regular sockets is shown between brackets.

Evaluation

Table 4: Roundtrip latency of SmartSockets (time in milliseconds).

	Source					
Target	DAS3-V	DAS3-D	Rockstar	Grid5000	Hiroshi	Desktop
DAS3-V		2.3 (2.3)	166 (166)	56	528	14 (14)
DAS3-D	2.3 (2.3)		167 (167)	57	533	15 (15)
Rockstar	166	167		205	590	195
Grid5000	56	57	205		524	50
Hiroshi	528	529	590	522		539
Desktop	14	15	190	43	522	

When applicable, the roundtrip latency of regular sockets is shown between brackets.

Table 5: Throughput of SmartSockets (in Mbit/second).

	Source					
Target	DAS3-V	DAS3-D	Rockstar	Grid5000	Hiroshi	Desktop
DAS3-V		182 (183)	2.6 (2.5)	2.5	0.25	0.65 (0.65)
DAS3-D	185 (186)		2.6 (2.5)	2.6	0.26	0.65 (0.65)
Rockstar	2.8	2.7		6.9	0.23	0.65
Grid5000	7.6	8.2	2.4		0.20	0.65
Hiroshi	0.73	0.73	0.70	0.73		0.61
Desktop	3.3	3.3	2.2	2.2	0.25	

When applicable, the throughput of regular sockets is shown between brackets.

vrije Universiteit

Future Work

- Alternative protocols
  - UDP instead of TCP
  - Parallel streams
  - Secure connections
- Dynamically switching approach
  - Find a better connection while the socket is in use!
- Hub network scalability

Summary

- Communicating on Grids is hard
  - Many connectivity problems occur
  - Takes a lot of work to find the problems and work around them
- SmartSockets reduces this into a single problem:
  - **How to set up a spanning tree of hubs**
- The rest is done automatically!

vrije Universiteit

vrije Universiteit

Questions ?

vrije Universiteit

Splicing

- Both sites simultaneously set up a connection
  - Connections meet in the middle
  - NAT requires port mapping prediction
- Results in direct connection
- Solves multi-firewall/NAT problems
  - Very sensitive to timing
    - Not guaranteed to work
  - Does not work for non-routed networks

vrije Universiteit

Other Features

- Reducing connectivity
  - Explicit configuration needed
  - Allows applications to simulate firewalls
  - Can be combined with traffic shaping
    - Simulate a Grid on your cluster!
- Visualization of connections
  - Gives insight into network problems

vrije Universiteit

Visualization

- SHOW EXAMPLE!

vrije Universiteit

Programming Interface

- Current (Java) implementation based on SocketFactory pattern
  - Also used for SSL connections
- Connection results in 'Socket'
  - Compatible with existing applications
- Main problem: Addressing!

- Easy to plug into applications
  - Provided that this model is used!

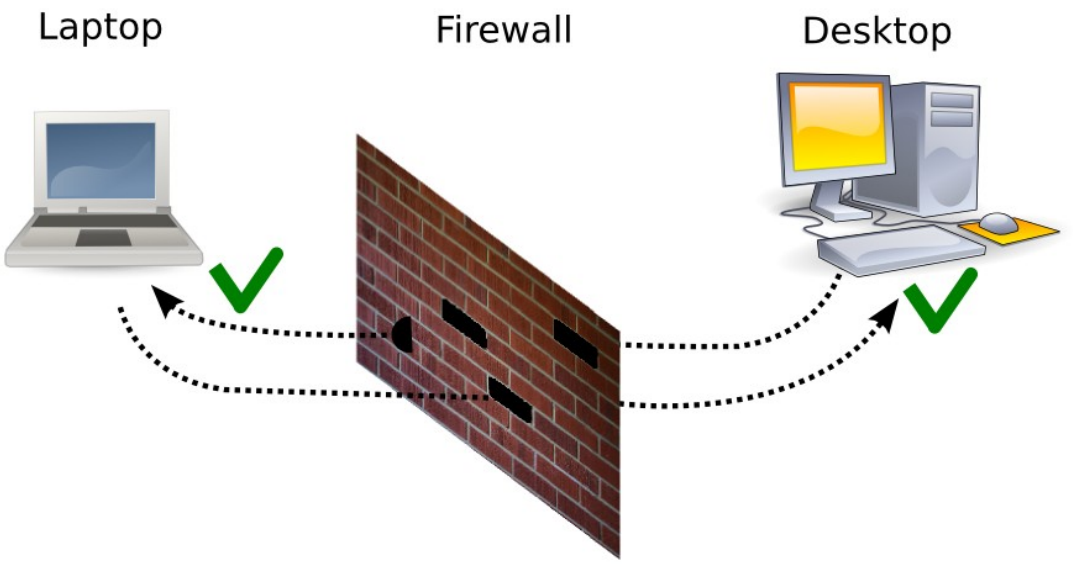
Multi Homing

- Which address should be used ?
  - Nodes can use both 193.x and 10.x addresses to connect to each other
    - 10.x is preferred
  - A machine outside the cluster can only use 193.x to connect to a node
  - How should the nodes advertise themselves ?

vrije Universiteit

Current Solution: Open port range

- Several ports are 'open'
  - Seen as a security hazard
  - How do you find the port range ?



ibis

vrije Universiteit

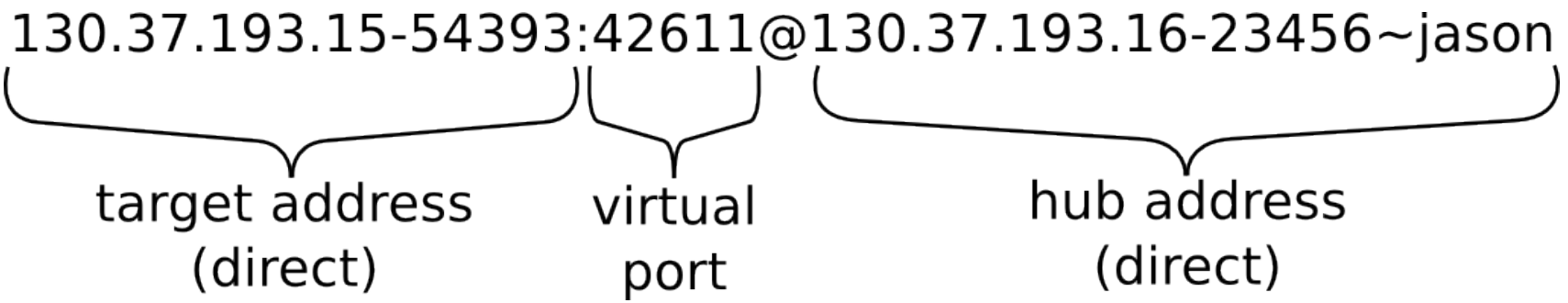
Current Solution: Port Forwarding

- Register ports at NAT device
  - Tell it where to forward incoming connections
- Problems:
  - How does the external machine know what IP/port combination to use ?
  - Not always supported
    - Usually only in consumer devices
    - Often 'switched off' (security problems)
  - Several different protocols

vrije Universiteit

Virtual Addressing

- Client addresses are extended with address of their hub
  - Needed for side-channel communication



Order Caching

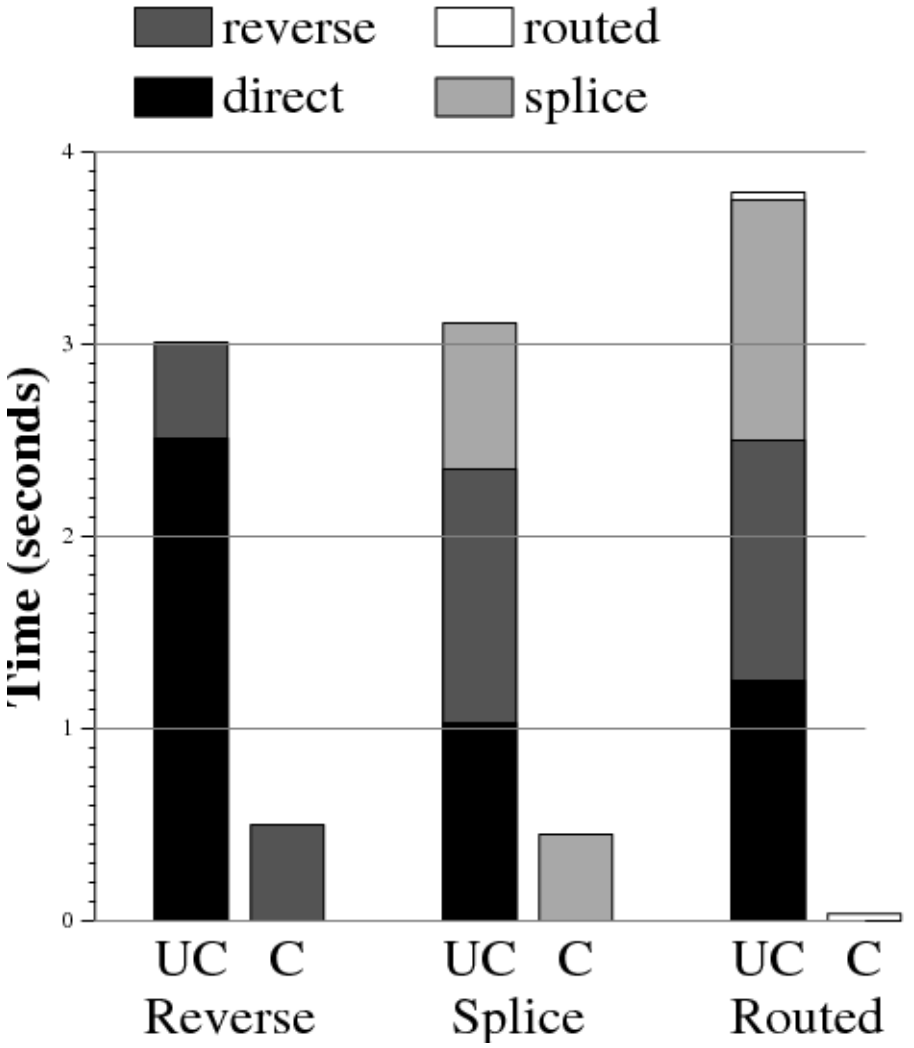
- By default the following order is used:

  direct, reverse, splice, routed

- Once a connection is established, the targets hub address and succesfull scheme are cached
- This scheme is tried first for the next connection to a target belonging to the same hub

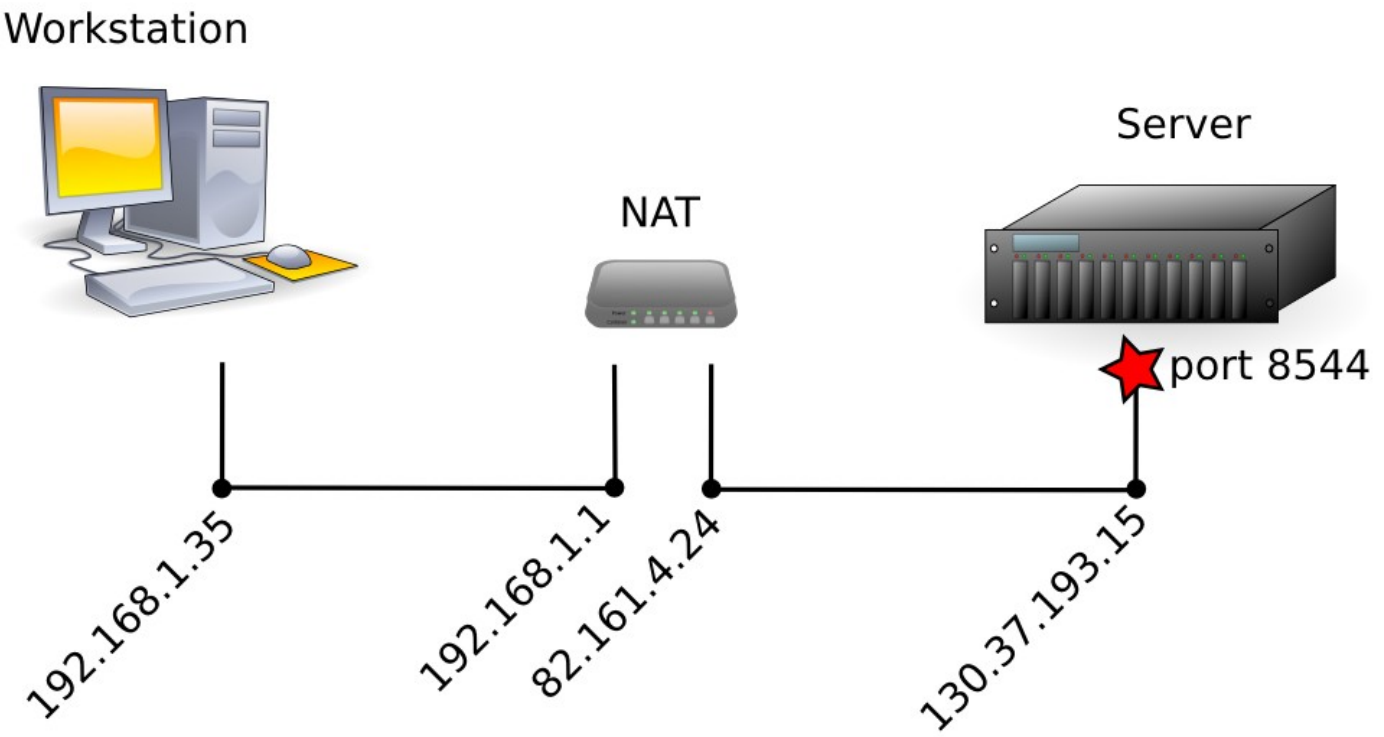
vrije Universiteit

Order Caching Results



vrije Universiteit

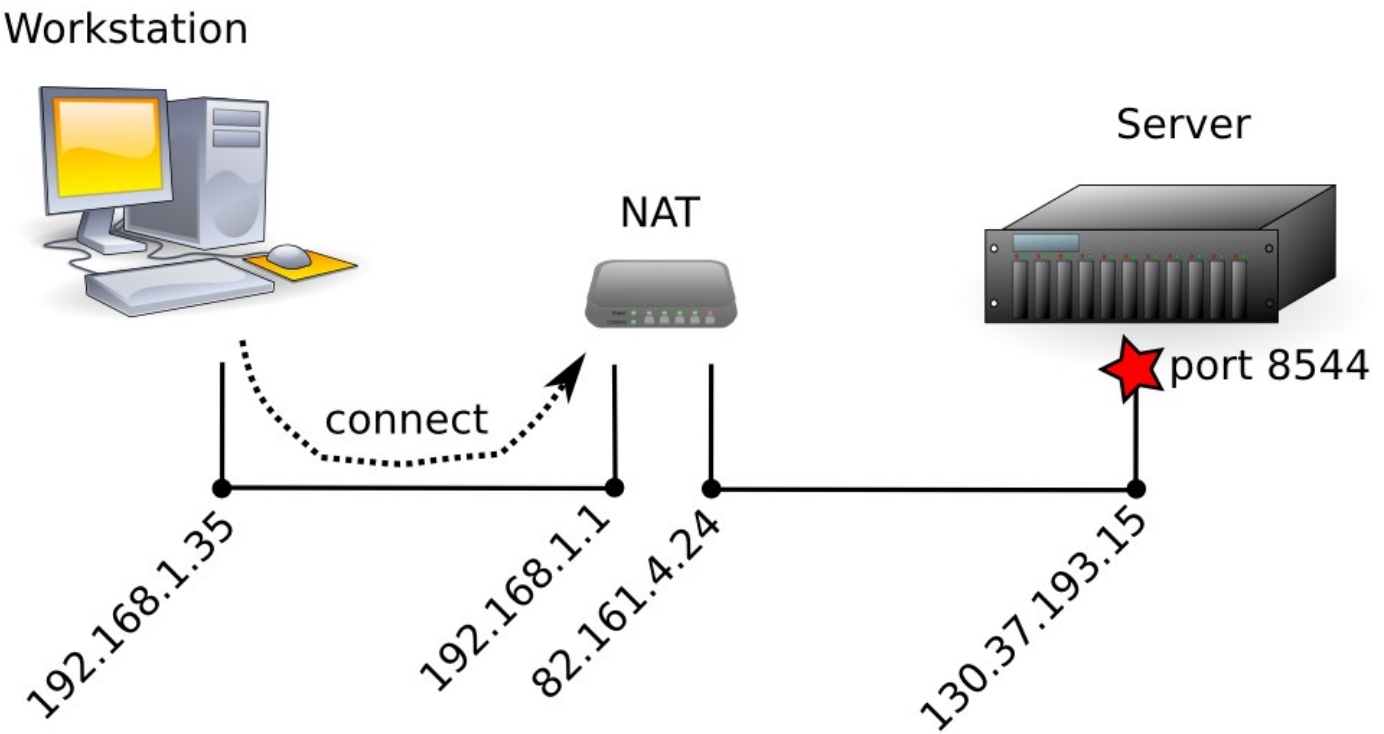
NAT: Outgoing connections



ibis

vrije Universiteit

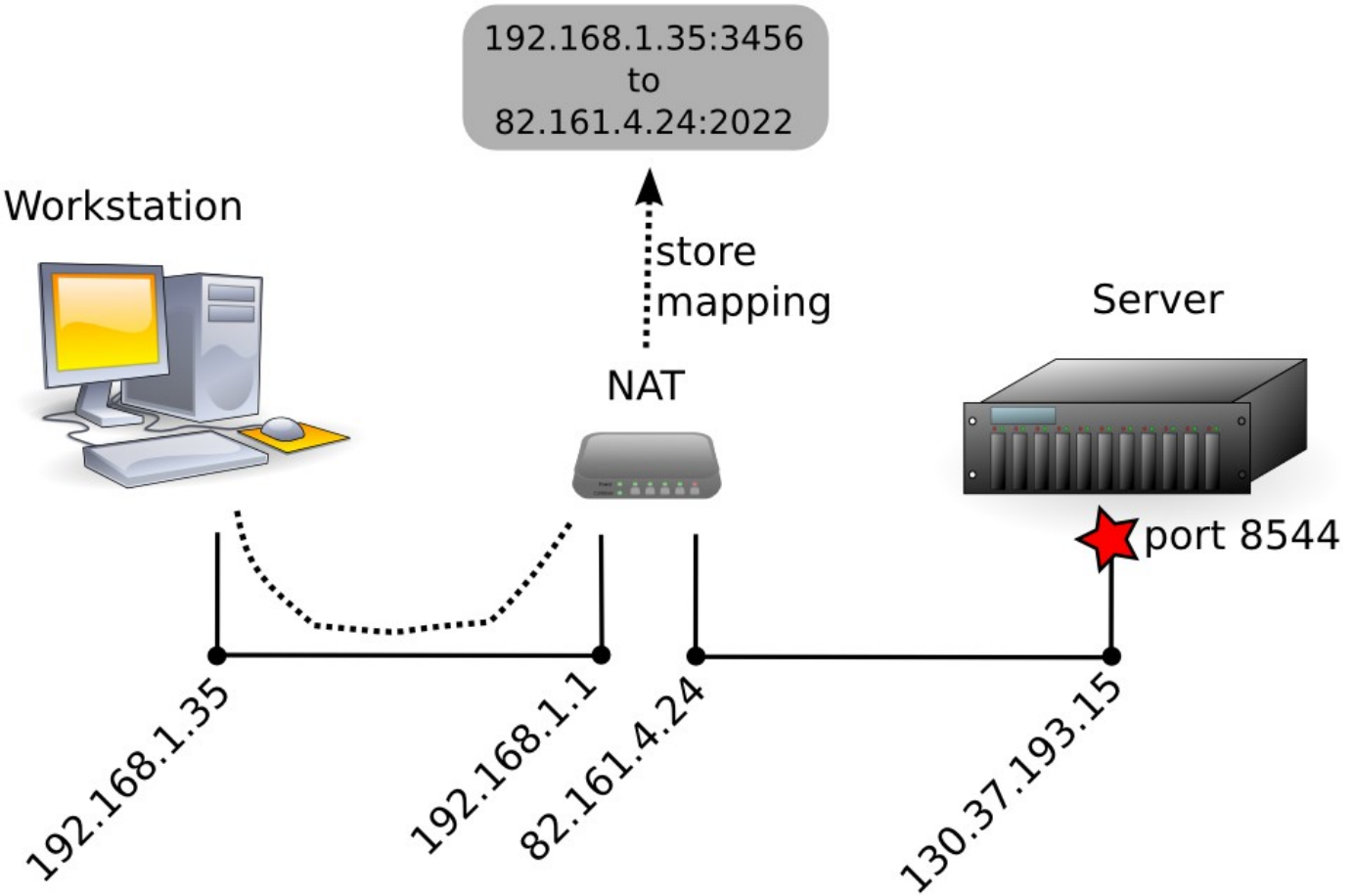
NAT: Outgoing connections



ibis

vrije Universiteit

NAT: Outgoing connections



ibis

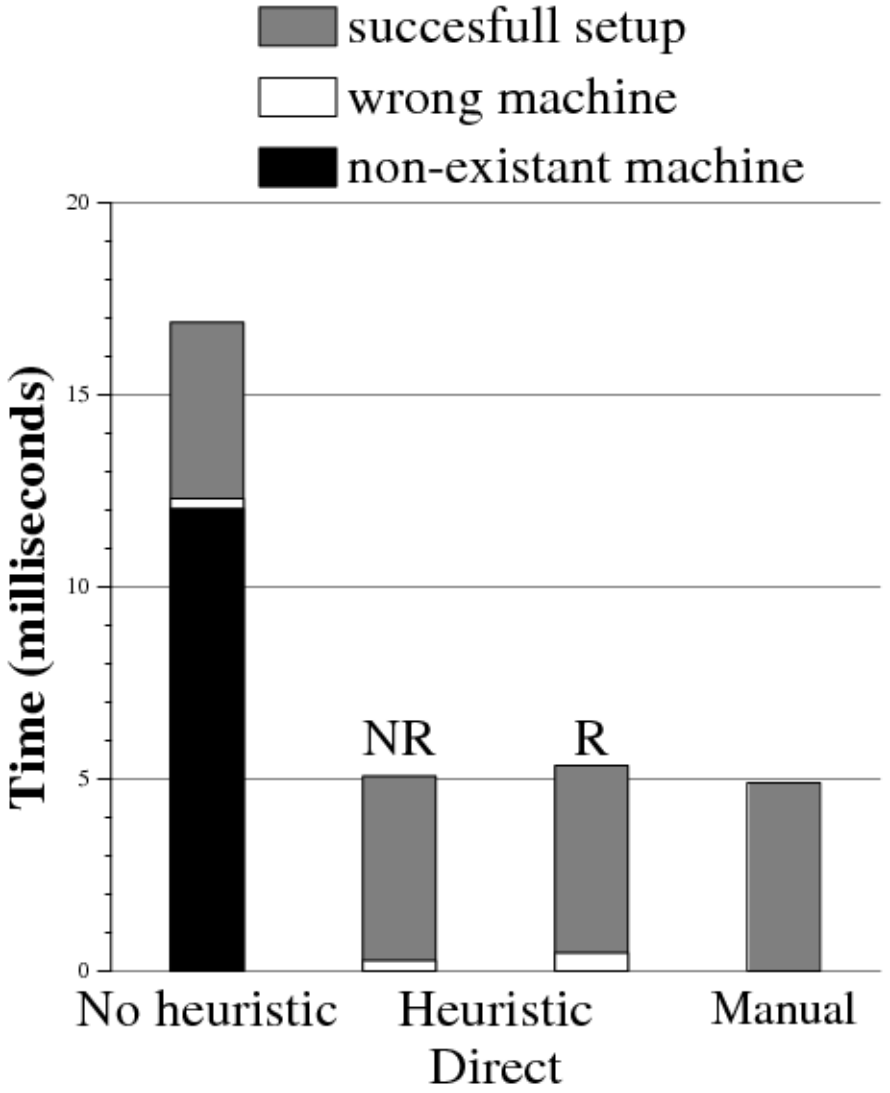
vrije Universiteit

Sorting Addresses

target address: 130.37.197.201 / 10.0.0.201

source address	result
130.37.193.15	130.37.197.201 / 10.0.0.201
130.37.193.15 / 192.168.1.15	130.37.197.201 / 10.0.0.201
192.168.1.15	130.37.197.201 / 10.0.0.201
10.0.0.15	10.0.0.201 / 130.37.197.201
130.37.193.15 / 10.0.0.15	10.0.0.201 / 130.37.197.201
192.168.1.15 / 10.0.0.15	10.0.0.201 / 130.37.197.201

Sorting Effect

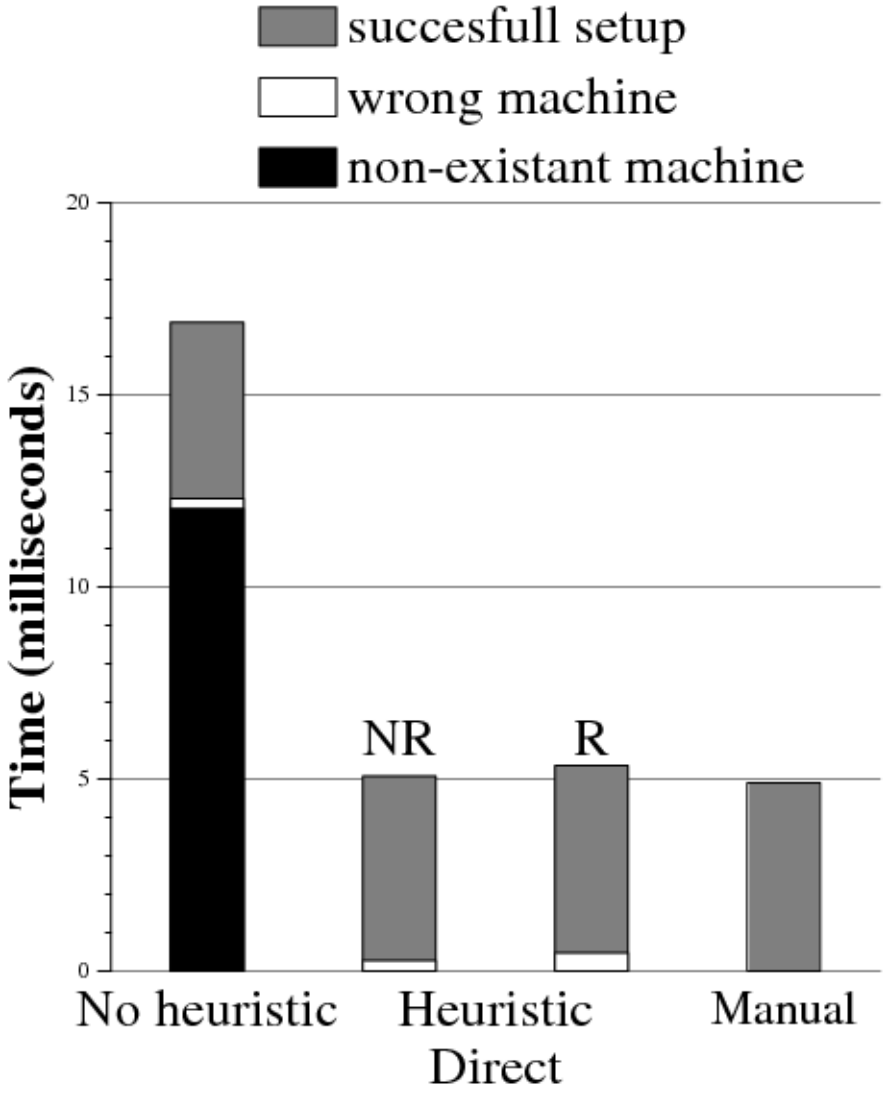


(DAS3-Delft to DAS3-VU)

ibis

vrije Universiteit

Sorting Effect



(DAS3-Delft to DAS3-VU)

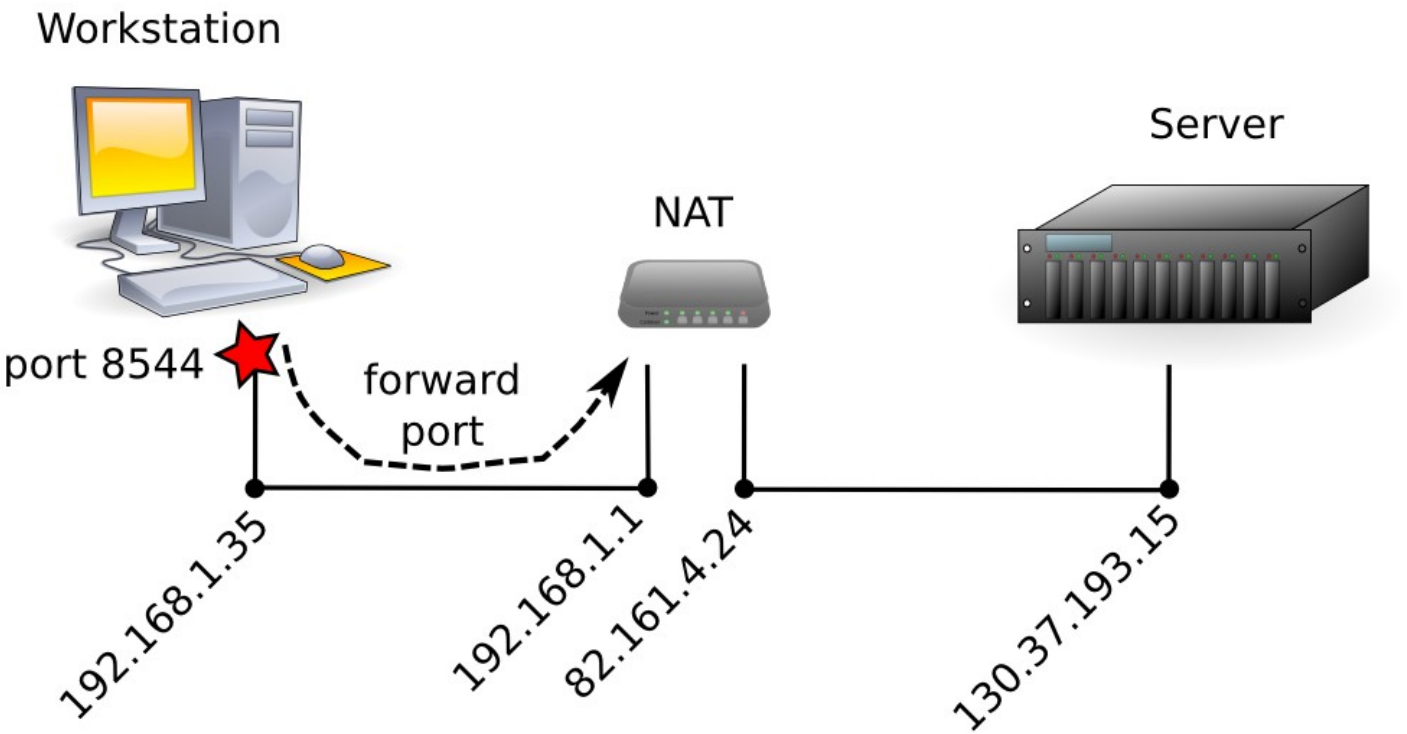
ibis

vrije Universiteit

Virtual Connection Layer

- Uses direct connection layer and hub network
- Connection setup schemes require cooperation between sites
  - (Direct)
  - Reverse
  - Splicing
  - Routed
- Mostly transparent to client!

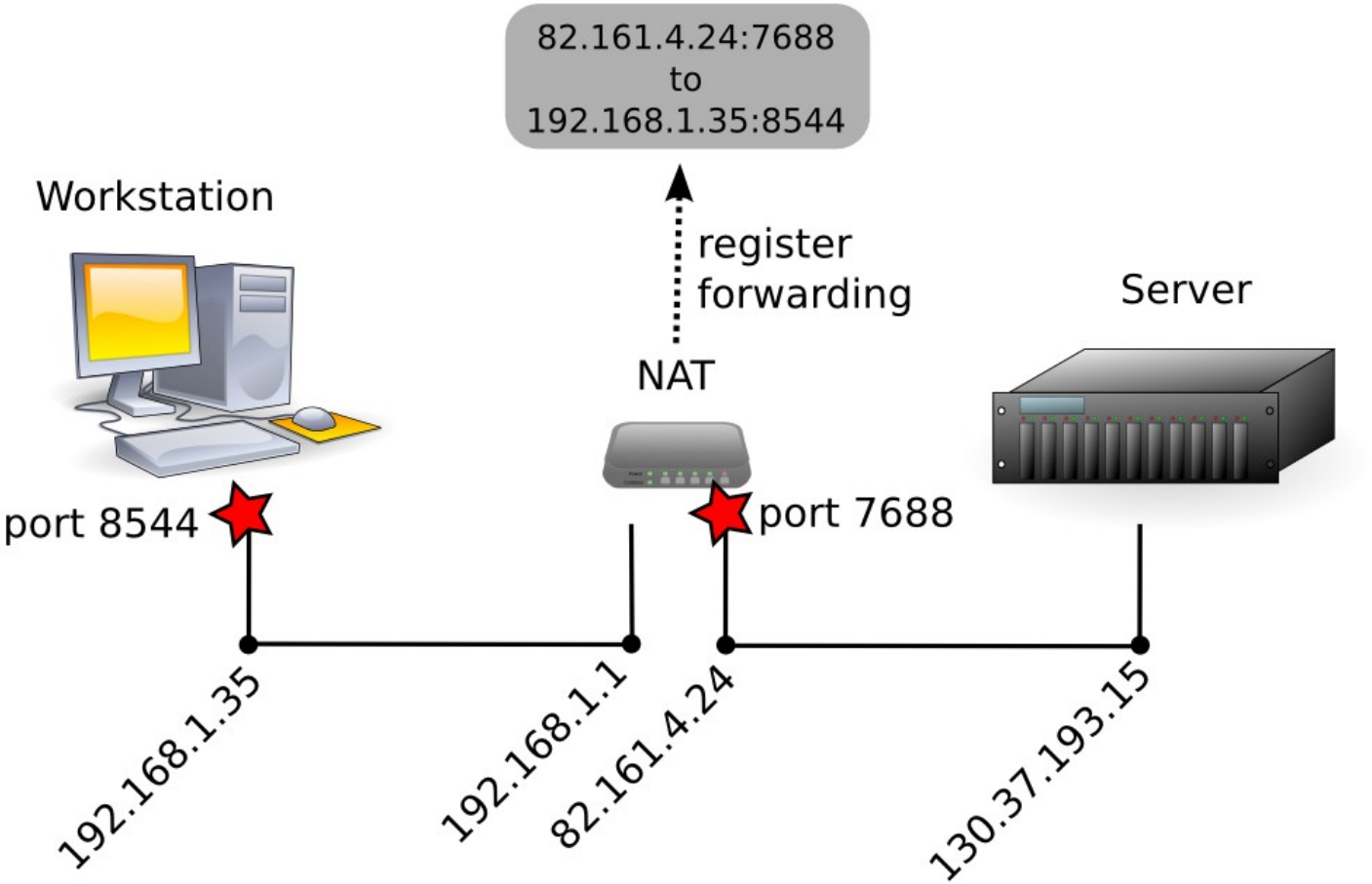
NAT: Port Forwarding



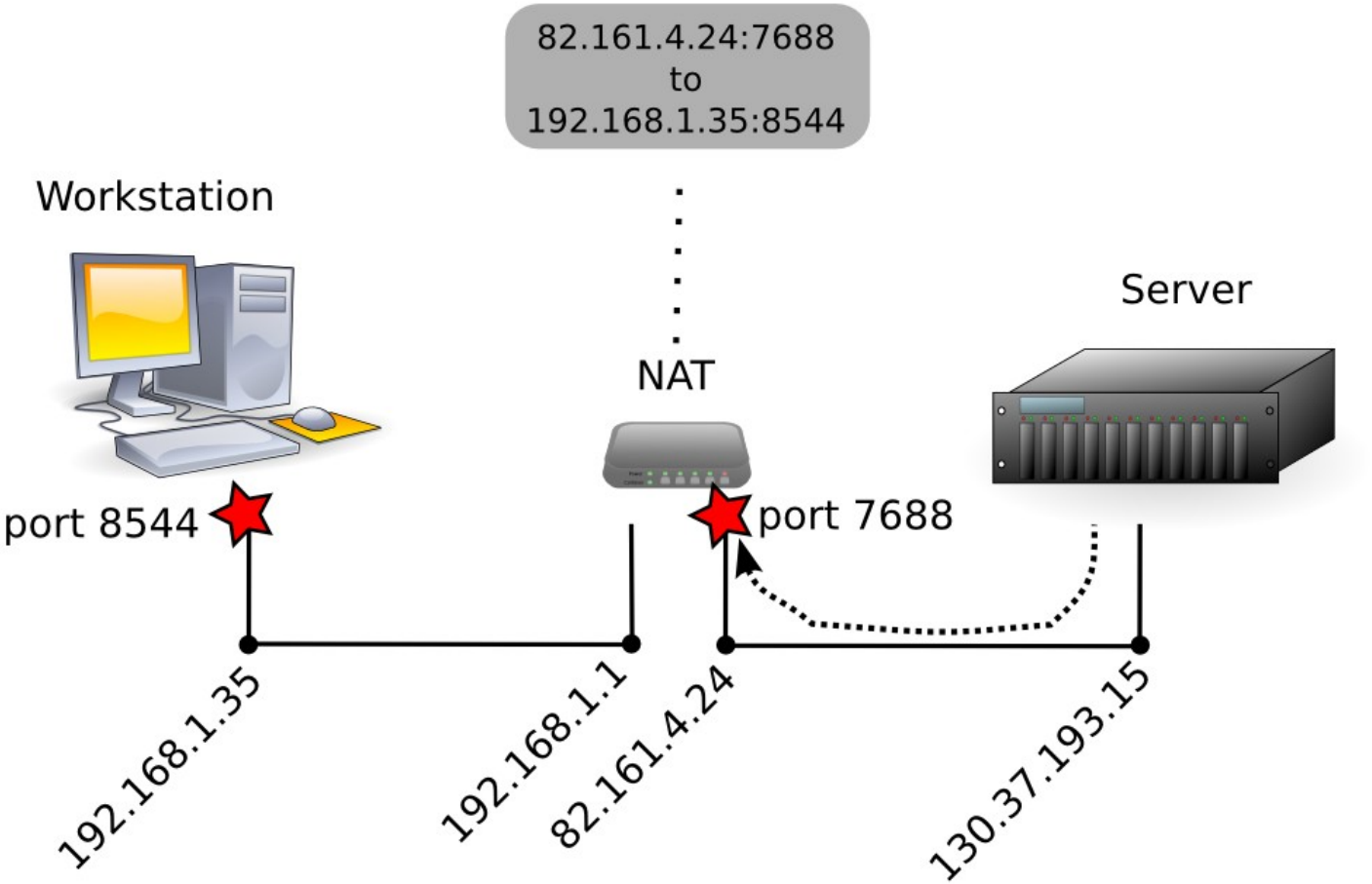
ibis

vrije Universiteit

NAT: Port Forwarding



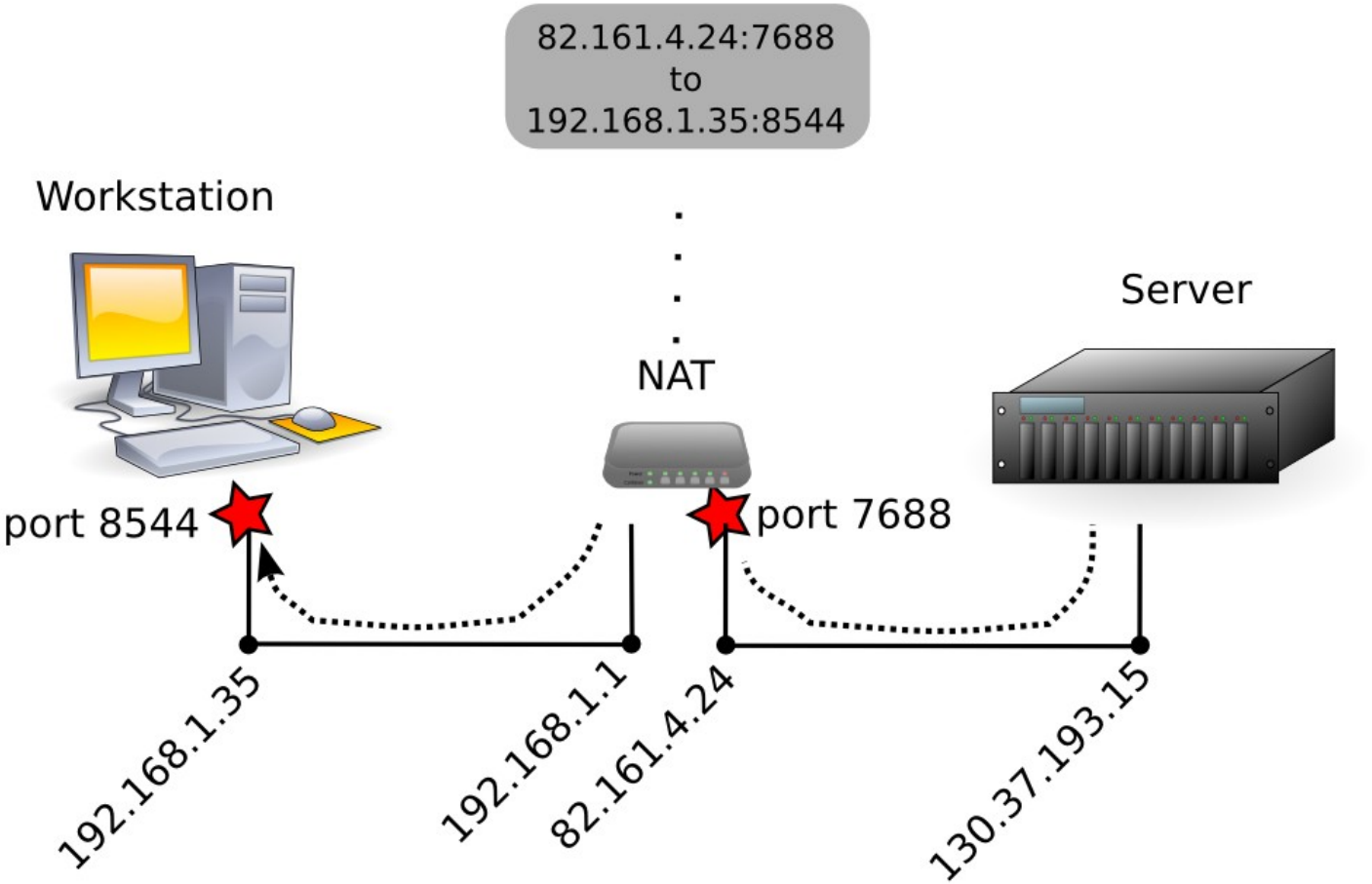
NAT: Port Forwarding



ibis

vrije Universiteit

NAT: Port Forwarding



Grid Computing
(our definition)

- To us a 'grid' is usually:
  - A collection of *clusters*
    - often a '*social collection*'
      - access often provided by (former) colleagues, friends, project partners, etc.
    - several administrative domains
      - difference in configurations, security settings and level of maintenance
  - May include desktop systems
    - Used for visualization, monitoring, steering

Heterogeneity

- Grids are heterogeneous
- We have developed languages and runtime systems that can handle the differences in processor and network speeds
- However, many applications assume that the participants can (directly) communicate

vrije Universiteit