

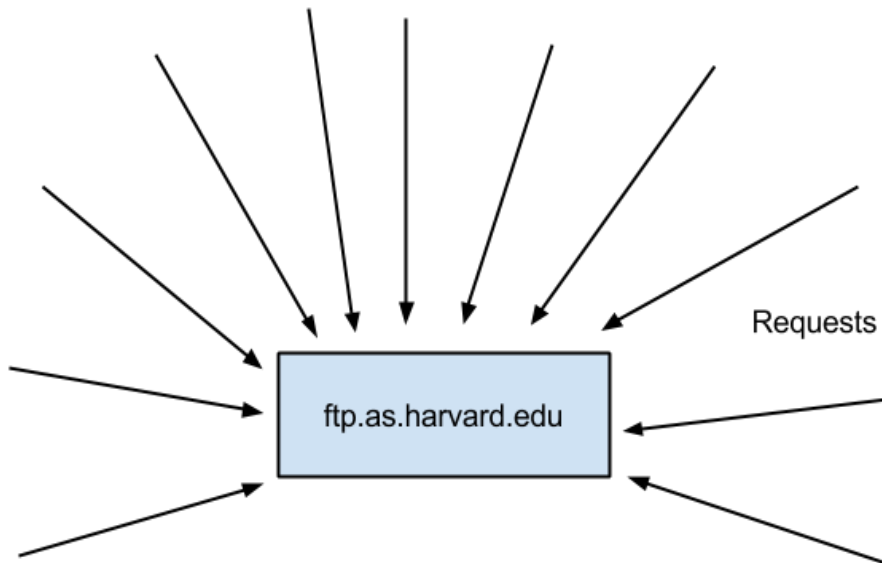
GC Data Distribution

Alternatives to FTP/HTTP

Source Node View

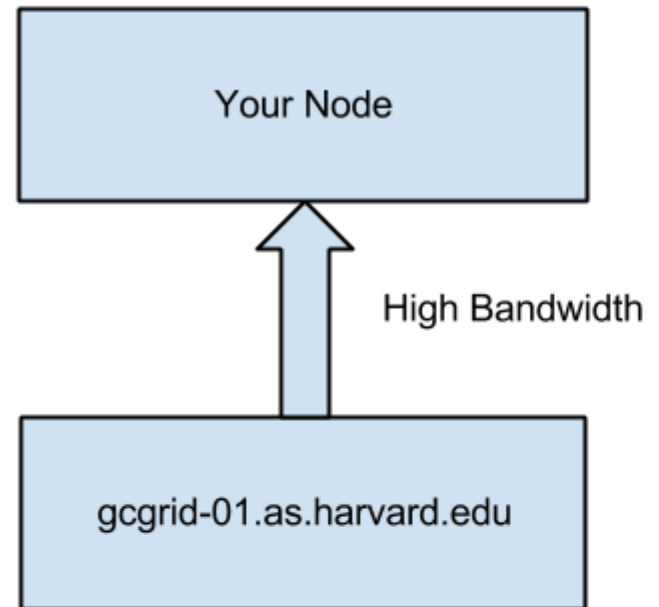
FTP

Shared Bandwidth

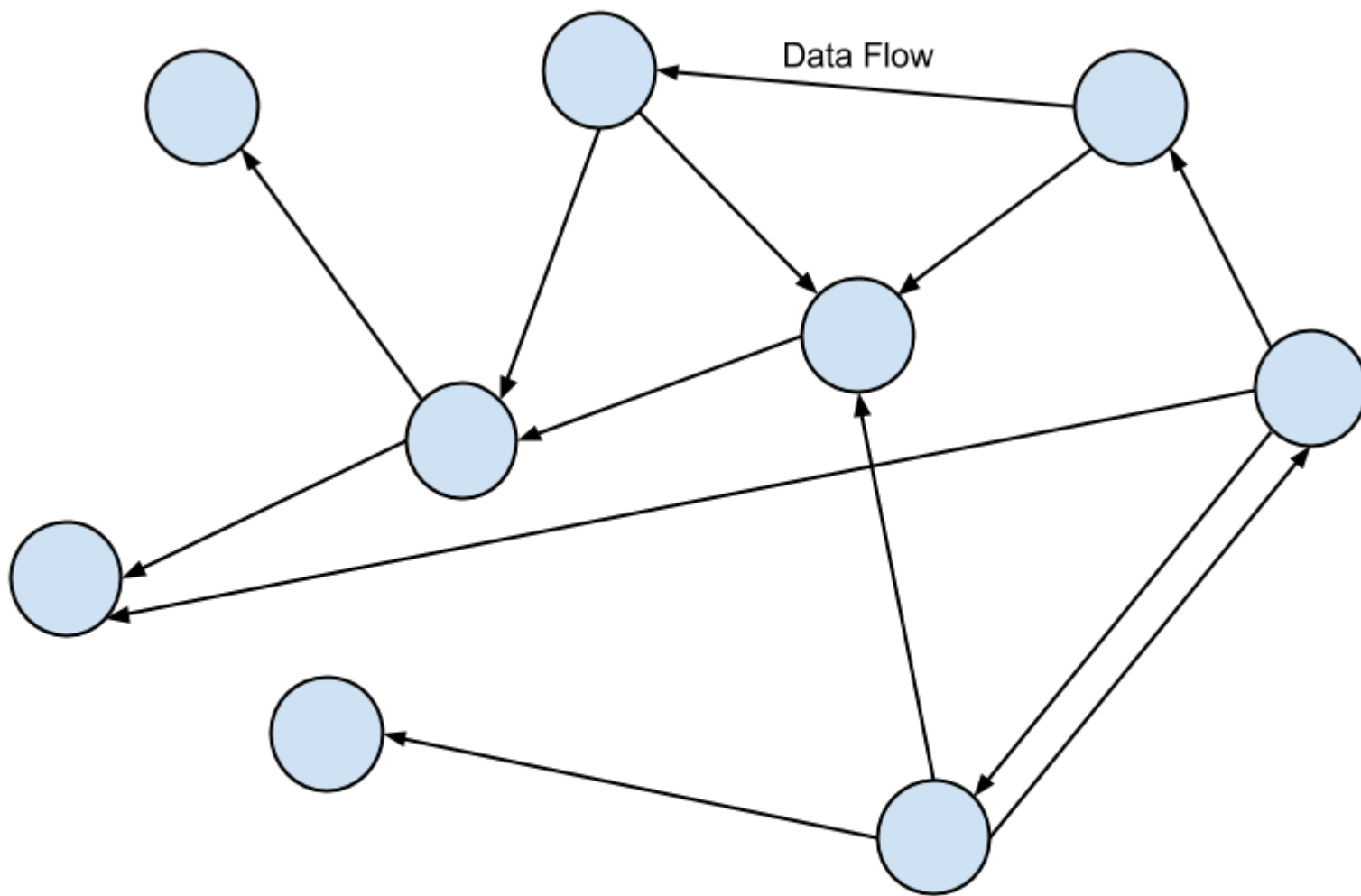


GCGrid

Dedicated Bandwidth



Grid Concept



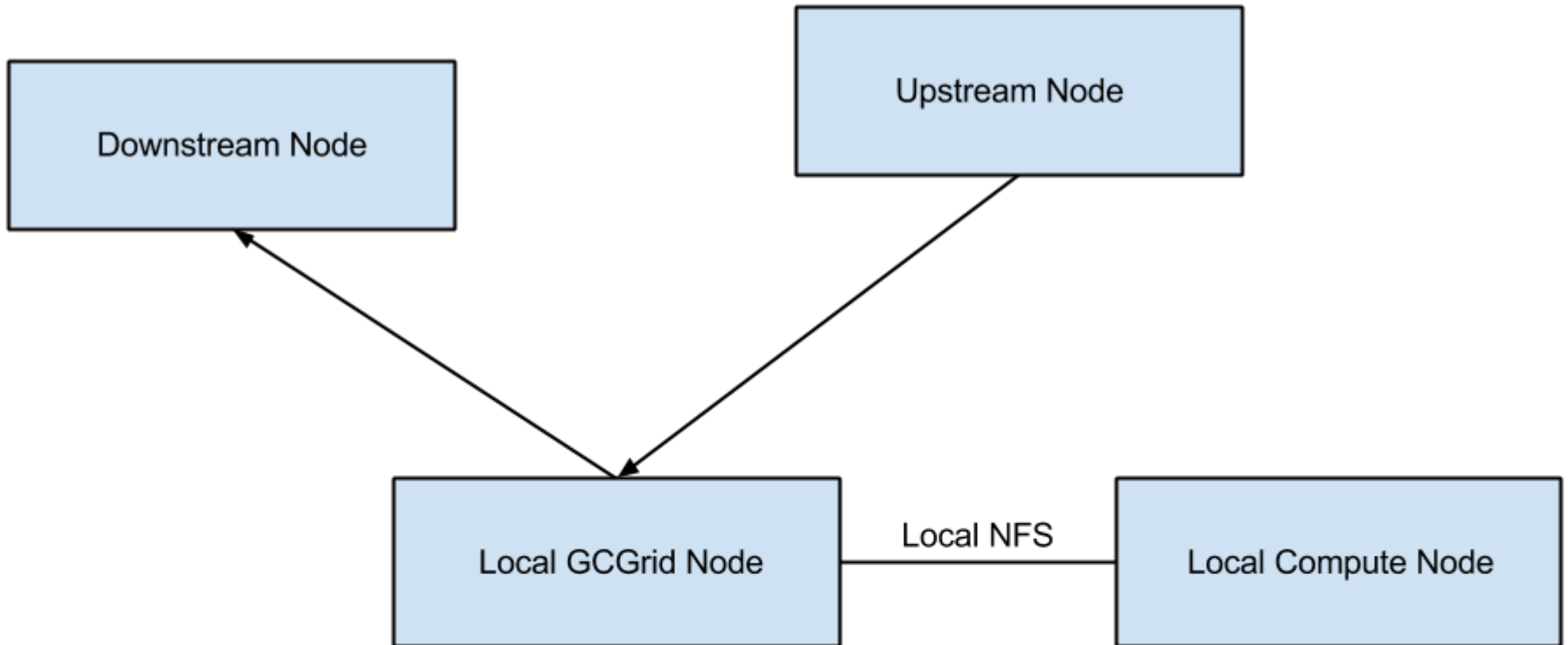
What is a GCGrid Node?

- A self-installing, self-configuring storage server on your local network
- An automated method for distributing data volumes to the GC community
- Receives data from at least one other node
- Supplies data to at least one other node
- Allows systems throughout your local domain to mount data volumes over your local high speed network

What a GCGrid node does not do

- Does not supply data to the entire internet (unless you choose to add that as a service)
- Does not grant privileged login access to a system at another institution

View from your institution



Central FTP/HTTP Service

Pro

- **Low Tech** (simple: web browser, ftp, wget, curl)
- **Self-service** (selective, last-minute access)
- **Free** (No equipment contribution required)

Con

- **20 year old technology**
- **Labor intensive**
- **Relies entirely on 1 or 2 sites**
- **Slow** (competition for limited bandwidth, interrupted transfers require restarts and monitoring)
- **Risks to data integrity** (corrupt/incomplete transfers, missed updates)

GEOS-Chem Grid Node

Pro

- Automated
- Fast (directly access local copy of data)
- Decentralized
- Maintains data integrity

Con

- Resources required (\$)
- Linux installation required
- Firewall change required
- Responsible for something beyond your research group

FTP/HTTP vs. GCGrid Node

FTP/HTTP Service

Cheaper and easier:

- Save money on equipment
- Avoid interacting with local IT staff

GCGrid Node

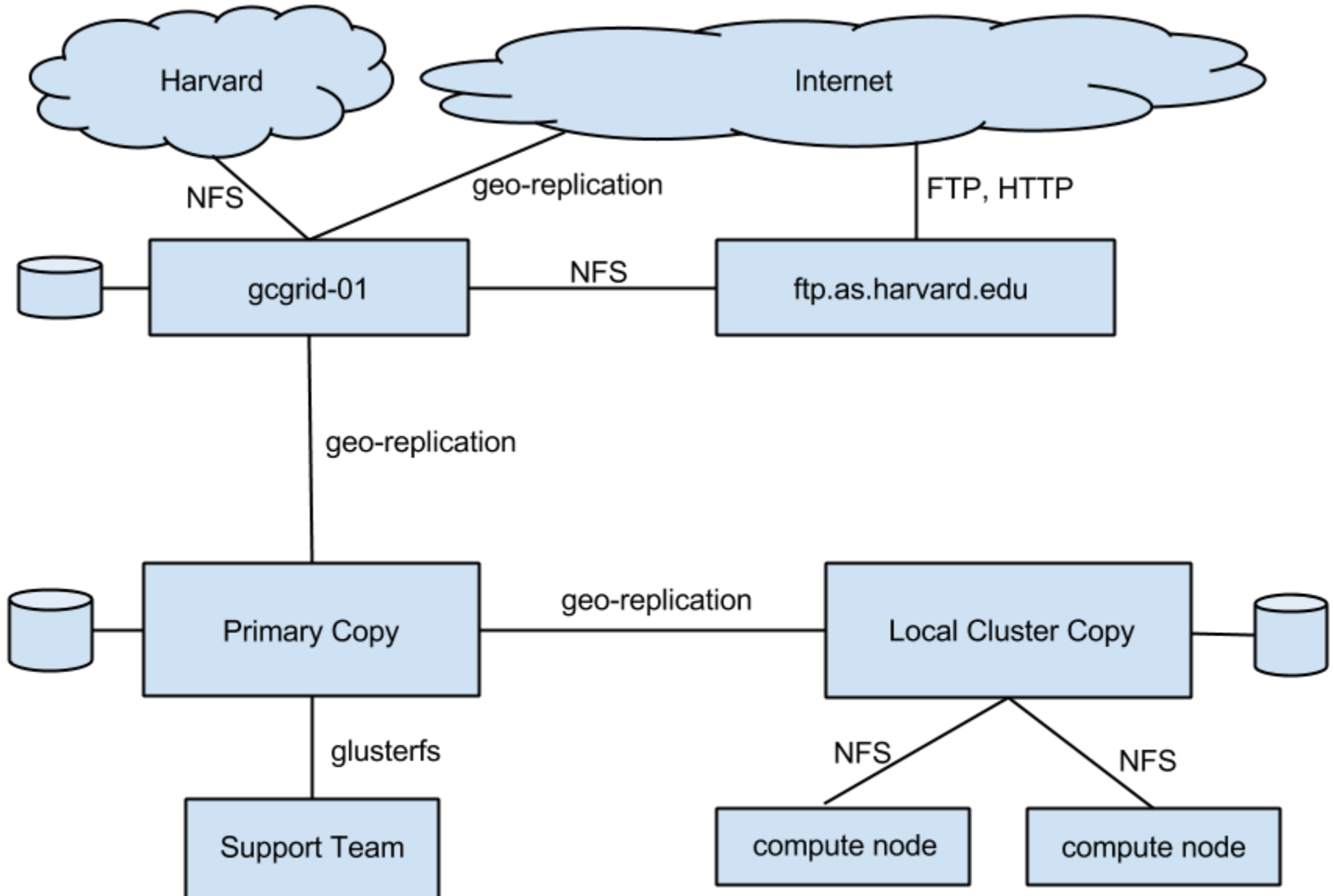
Improves productivity:

- Save human labor (more valuable than equipment?)
- Faster
- Safer using latest version of data

GCGrid Node Lite:

We can also replicate data using periodic rsync to institutions willing to serve data to the entire internet using anonymous FTP or HTTP.

Harvard ACMG Nodes



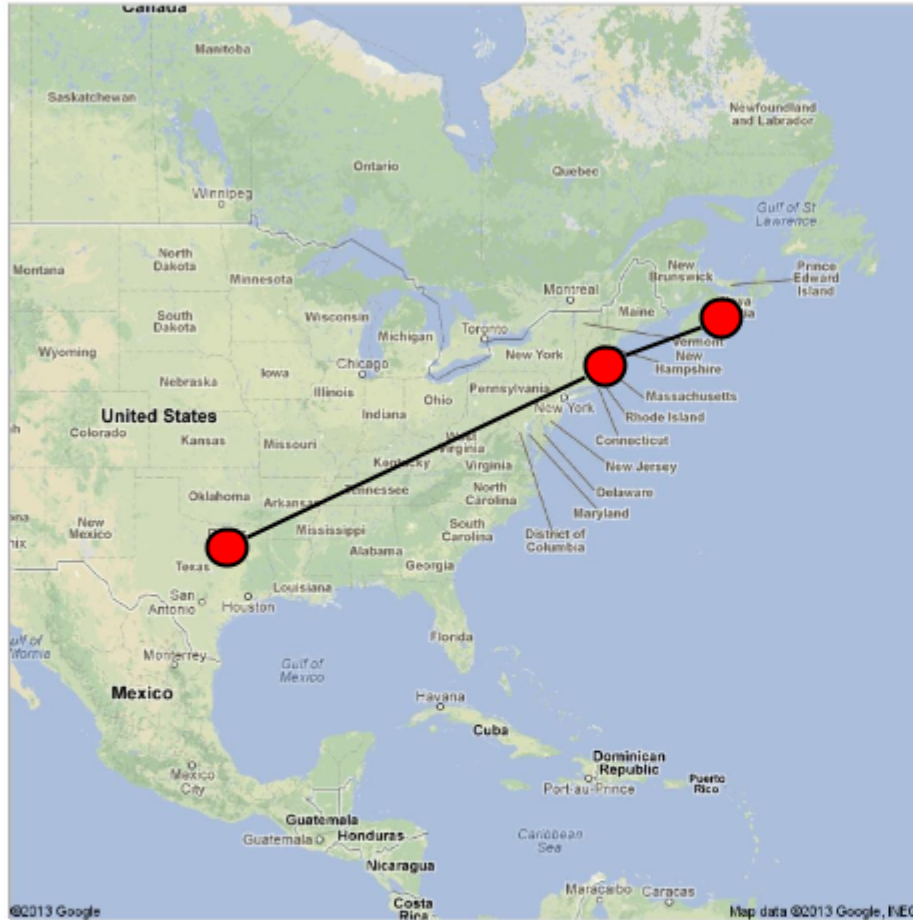
GCGrid Today



Get Google Maps on your phone



Text the word "GMAPS" to 466453



Future?



System Requirements

- Dedicated Linux system running CentOS 6, Fedora 16, or higher
- 8GB physical memory (12GB recommended)
- 4 Cores (8 recommended)
- 2TB minimum. 10TB block storage (16TB recommended) for all data except 2x2.5 MERRA. 24TB for all data including 2x2.5 MERRA (32TB recommended)
- 100Mbps ethernet connection (1Gbps recommended)
- Static IP address (routable IPv4) and name (i.e. don't change)
- Static mount point for the storage (don't change)

Systems with fewer resources can act as GCGrid nodes if they do not replicate all of the data volumes. Storage less than 2TB is not supported.

Reference: http://wiki.seas.harvard.edu/geos-chem/index.php/Building_a_GCGrid_node

Contact: jhy@as.harvard.edu