Cloud-hosted Data Transfer & Optimization:

Stork for the Cloud

Tevfik Kosar

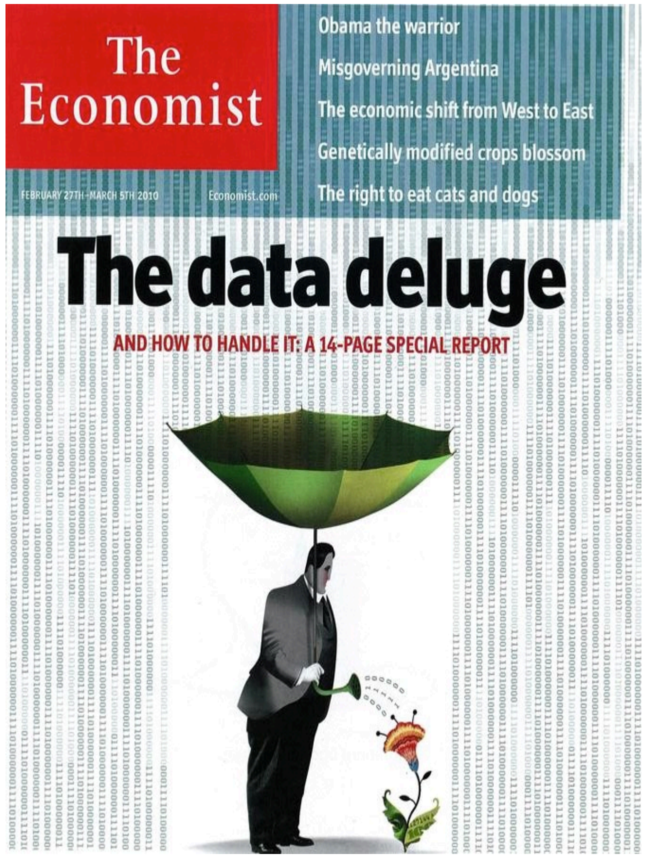
University at Buffalo (SUNY)

May 2, 2012

Condor Week, Madison, WI

Big Data

Science



The Economist

Obama the warrior

Misgoverning Argentina

The economic shift from West to East

Genetically modified crops blossom

The right to eat cats and dogs

FEBRUARY 27TH–MARCH 5TH 2010

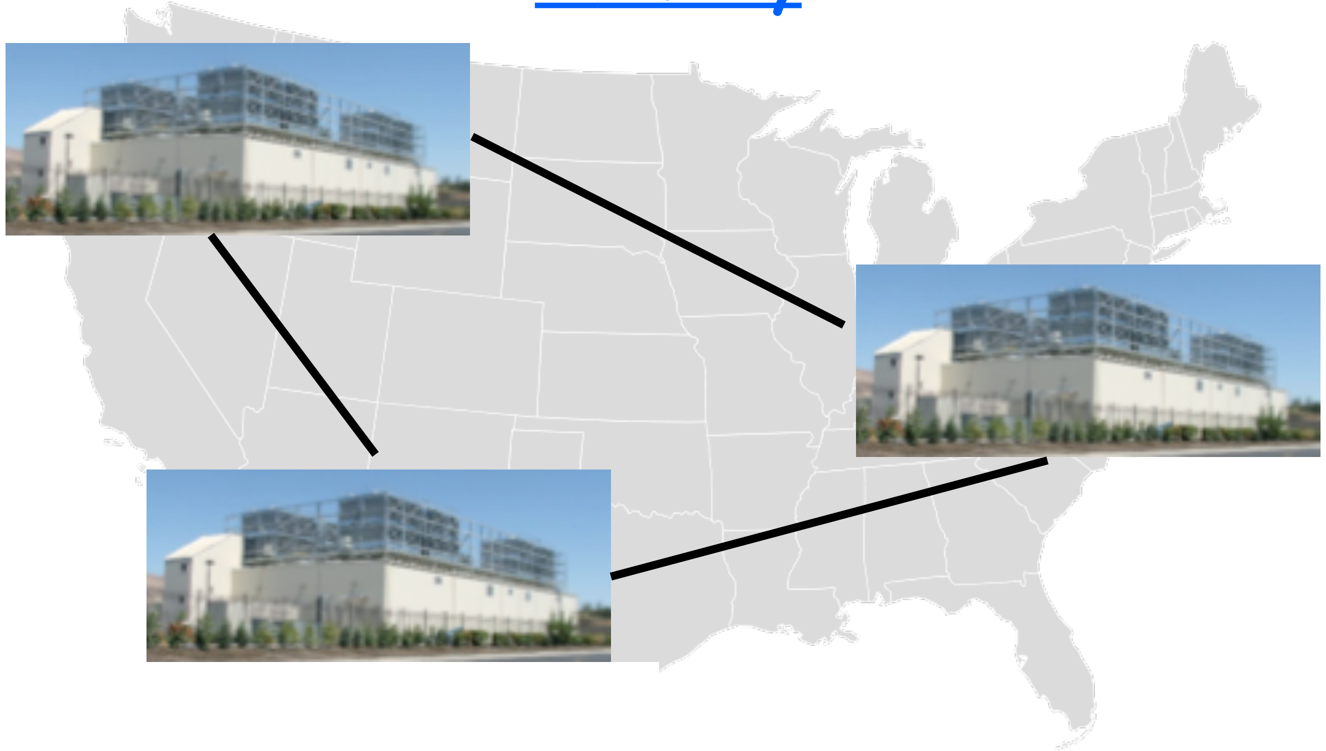
Economist.com

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

- 1 PB is now considered "small" for many science applications today
- For most, their data is distributed across several sites

Industry

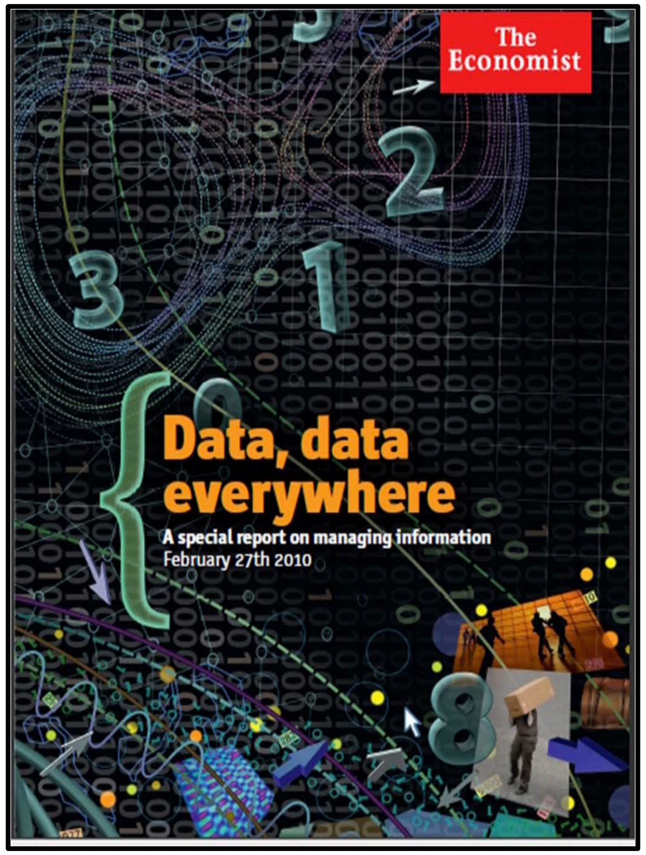


A survey among 106 organizations operating two or more data centers:

- 77% run replication among three or more sites
- 50% has more than 1 PB in their primary data center

Big Data

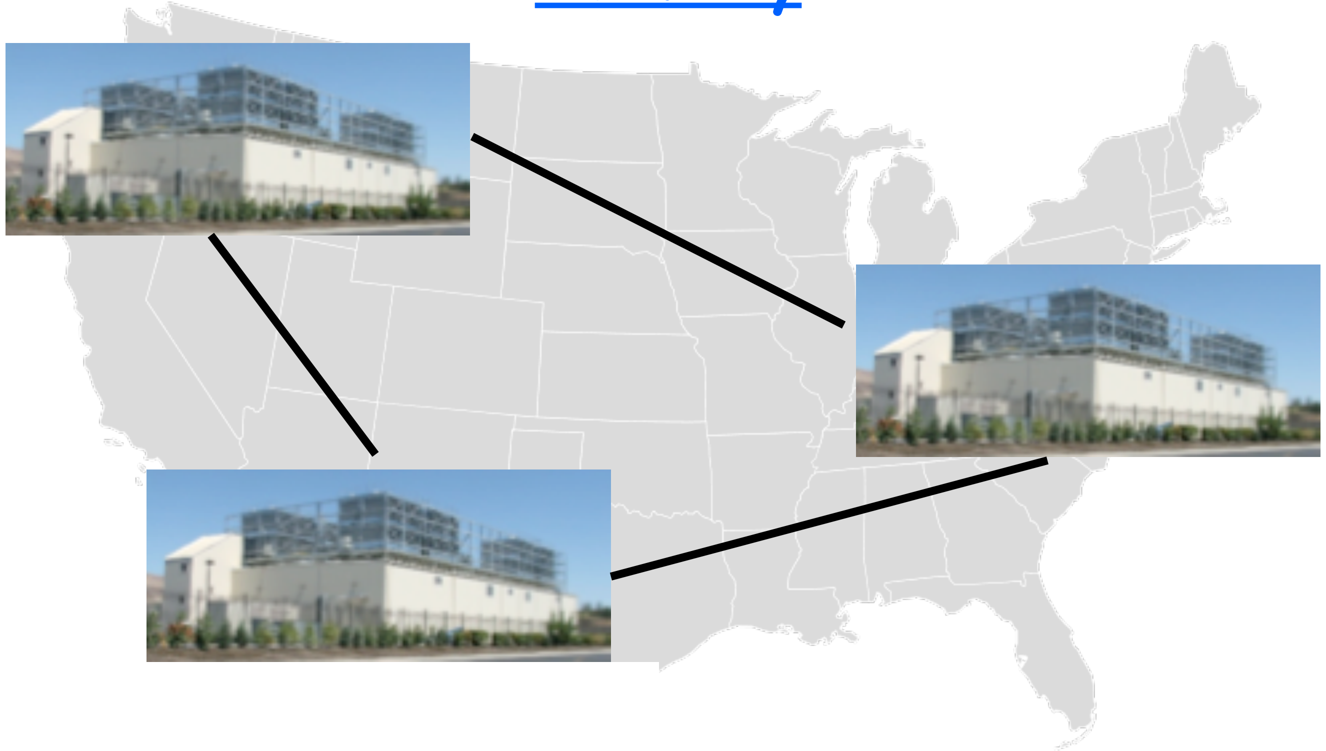
Science



- 1 PB is now considered "small" for many science applications today

- For most, their data is distributed across several sites

Industry



A survey among 106 organizations operating two or more data centers:

- 77% run replication among three or more sites

- 50% has more than 1 PB in their primary data center

Best Way to Move Big Data?

Best Way to Move Big Data?

- Sending **1 PB** of data over 10 Gbps link would take **nine days** (assuming 100% efficiency) -- too optimistic!

Best Way to Move Big Data?

- Sending **1 PB** of data over 10 Gbps link would take **nine days** (assuming 100% efficiency) -- too optimistic!
- Sending **1 TB** Forensics dataset from Boston to Amazon S3 cost $100 and took **several weeks** [Garfinkel 2007]

Best Way to Move Big Data?

- Sending **1 PB** of data over 10 Gbps link would take **nine days** (assuming 100% efficiency) -- too optimistic!
- Sending **1 TB** Forensics dataset from Boston to Amazon S3 cost $100 and took **several weeks** [Garfinkel 2007]
- Visualization scientists at LANL dumping data to tapes and sending them to Sandia Lab via **Fedex** [Feng 2003]

Best Way to Move Big Data?

- Sending **1 PB** of data over 10 Gbps link would take **nine days** (assuming 100% efficiency) -- too optimistic!
- Sending **1 TB** Forensics dataset from Boston to Amazon S3 cost $100 and took **several weeks** [Garfinkel 2007]
- Visualization scientists at LANL dumping data to tapes and sending them to Sandia Lab via **Fedex** [Feng 2003]
- Collaborators have the option of moving their data into disks, and sending them as packages through **UPS or FedEx** [Cho et al 2011].

Best Way to Move Big Data?

- Sending **1 PB** of data over 10 Gbps link would take **nine days** (assuming 100% efficiency) -- too optimistic!
- Sending **1 TB** Forensics dataset from Boston to Amazon S3 cost $100 and took **several weeks** [Garfinkel 2007]
- Visualization scientists at LANL dumping data to tapes and sending them to Sandia Lab via **Fedex** [Feng 2003]
- Collaborators have the option of moving their data into disks, and sending them as packages through **UPS or FedEx** [Cho et al 2011].
- Will **100 Gbps** networks change anything?

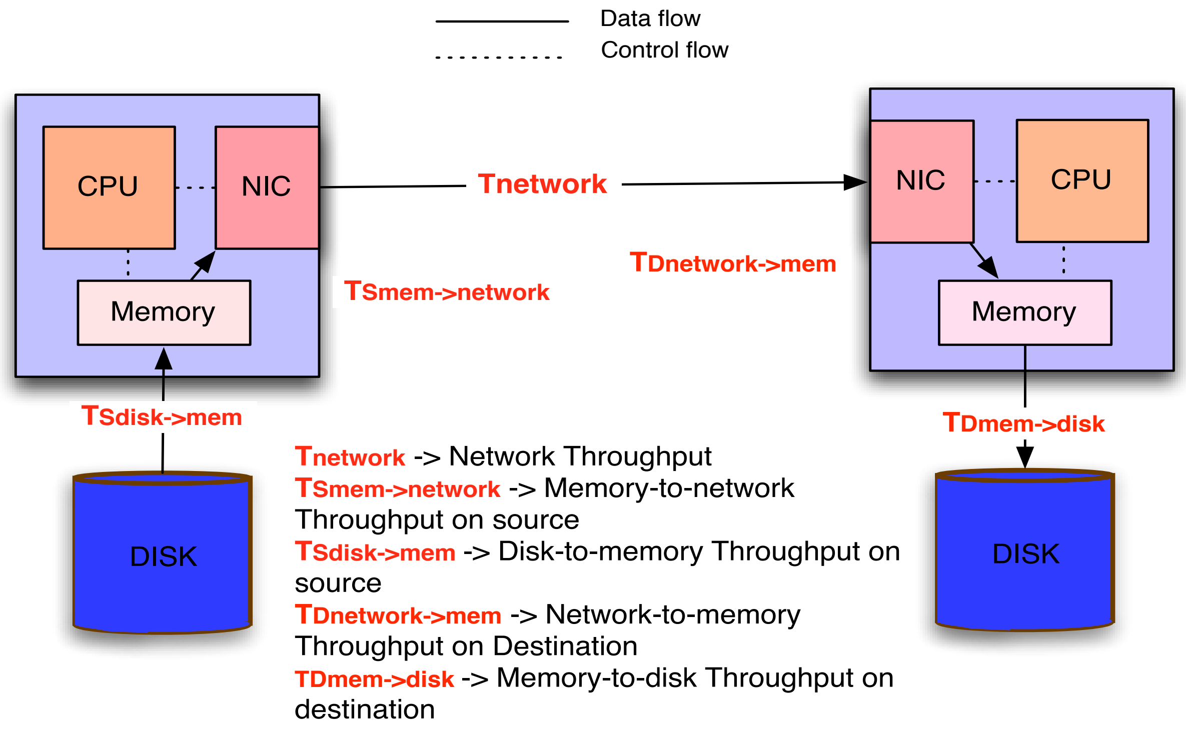
Stork Data Scheduler

- Implements state-of-the art models and algorithms for data scheduling & optimization
- Started as part of the Condor Project (was my PhD work)
- Currently developed at University at Buffalo and funded by NSF (CAREER, STCI, CiC)
- Based on the Condor code, uses Condor libraries (DaemonCore, ClassAds)
- Compatible with Condor products (i.e. DAGMan)
- .....

Stork Data Scheduler

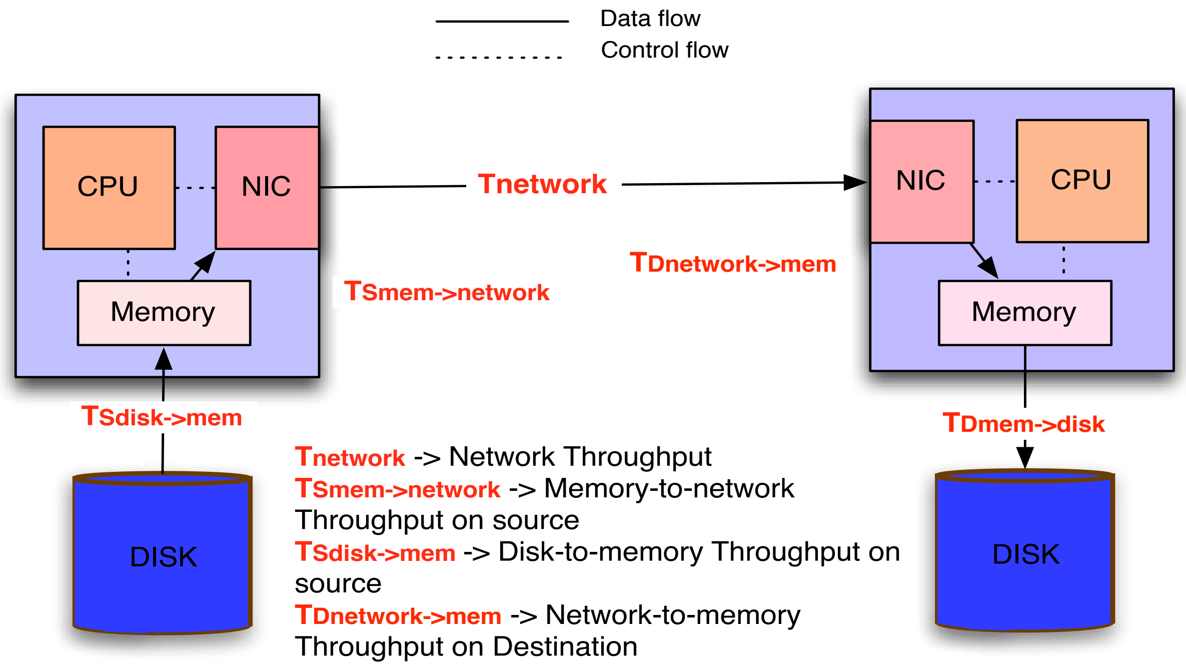
- .....
- Built & tested on Condor NMI (Metronome)
- Supports more than 20 platforms
- Futures include:
  - support for multiple transfer protocols
  - dynamic protocol tuning & optimization
  - end-to-end throughput prediction services
  - data aggregation & connection caching
  - early error detection and classification & recovery

End-to-end Problem



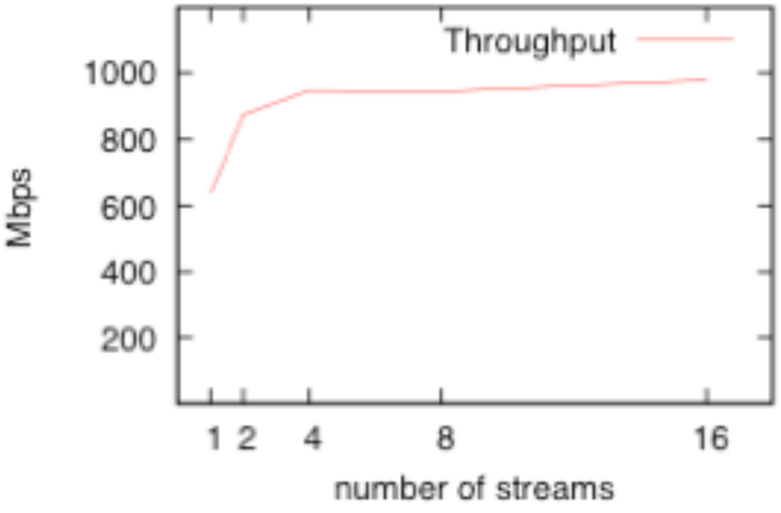
Tnetwork -> Network Throughput
TSmem->network -> Memory-to-network Throughput on source
TSdisk->mem -> Disk-to-memory Throughput on source
TDnetwork->mem -> Network-to-memory Throughput on Destination
TDmem->disk -> Memory-to-disk Throughput on destination

End-to-end Problem

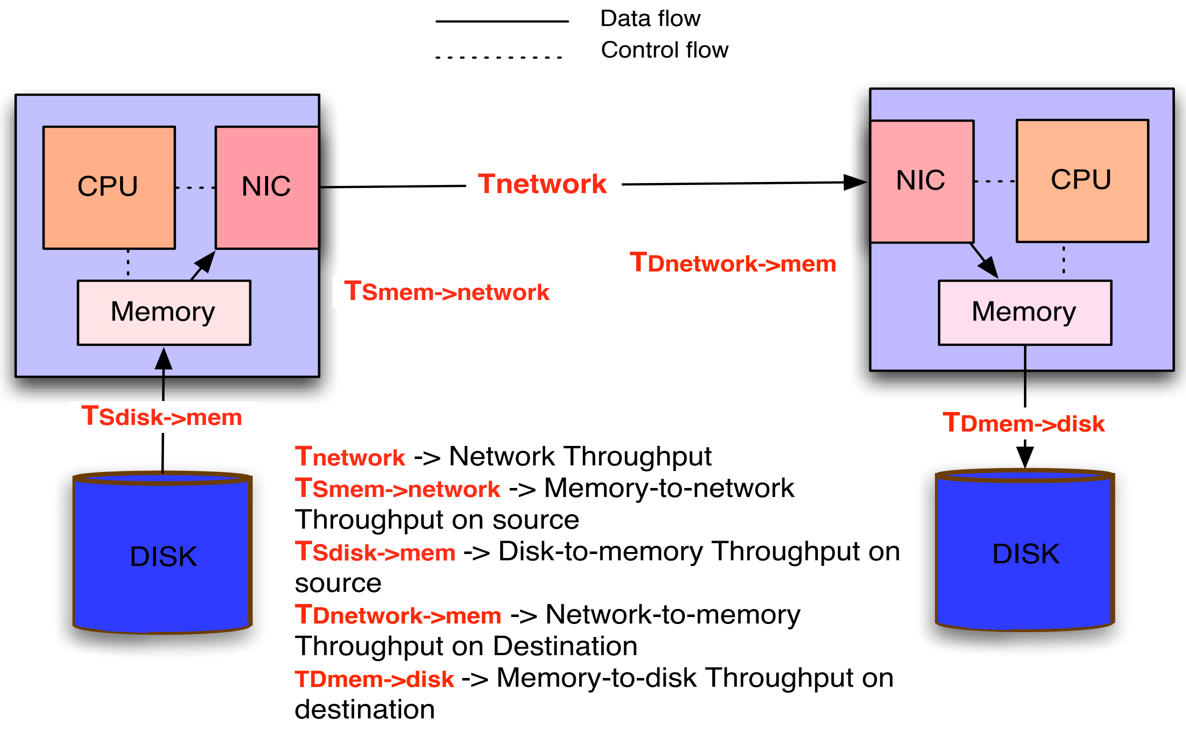


Tnetwork -> Network Throughput
TSmem->network -> Memory-to-network Throughput on source
TSdisk->mem -> Disk-to-memory Throughput on source
TDnetwork->mem -> Network-to-memory Throughput on Destination
TDmem->disk -> Memory-to-disk Throughput on destination

protocol tuning

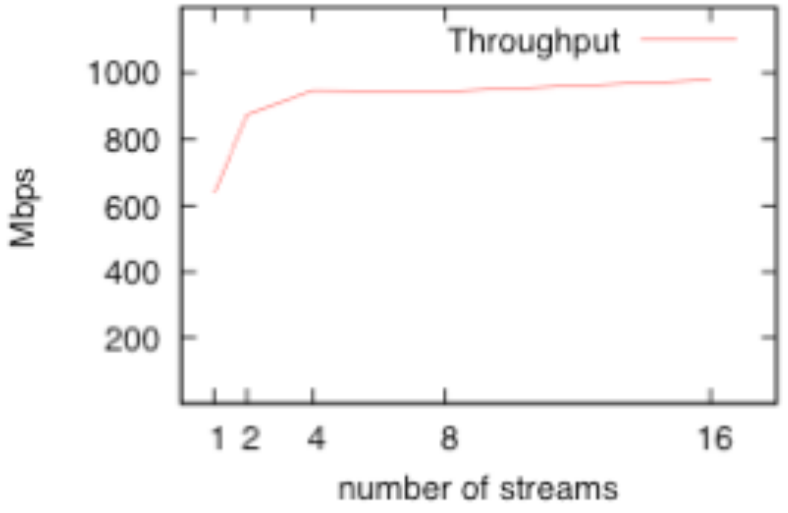


End-to-end Problem

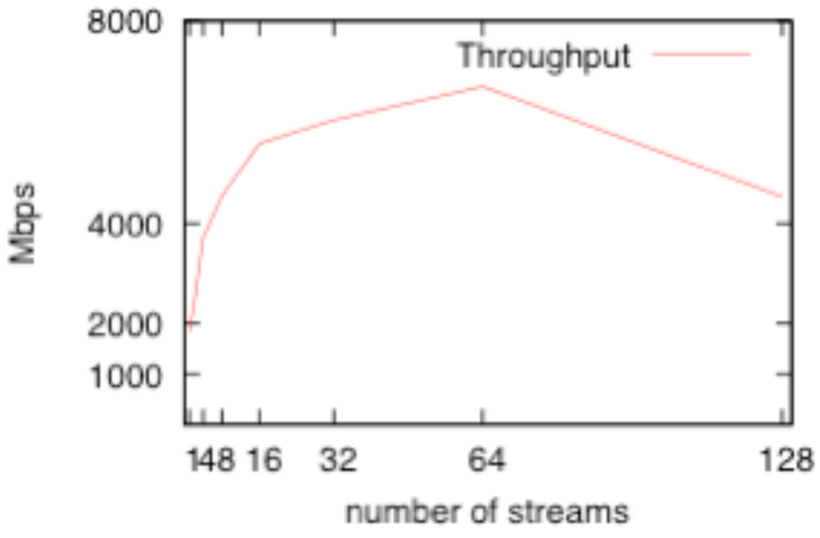


Tnetwork -> Network Throughput
TSmem->network -> Memory-to-network Throughput on source
TSdisk->mem -> Disk-to-memory Throughput on source
TDnetwork->mem -> Network-to-memory Throughput on Destination
TDmem->disk -> Memory-to-disk Throughput on destination

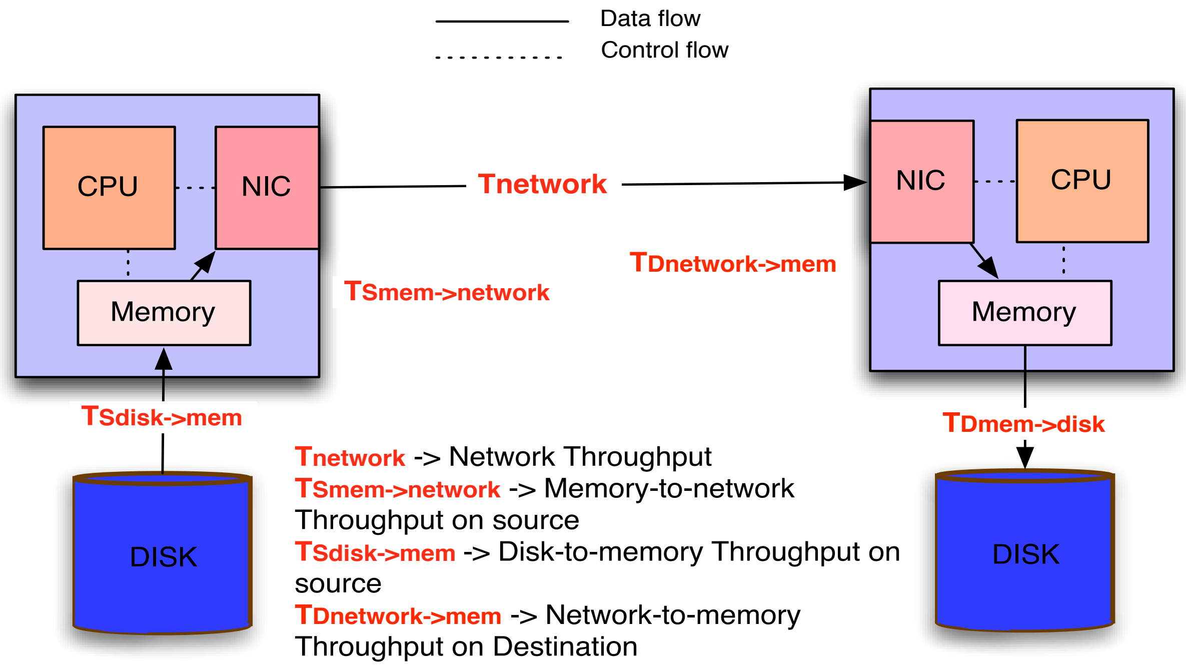
protocol tuning



disk I/O optimization

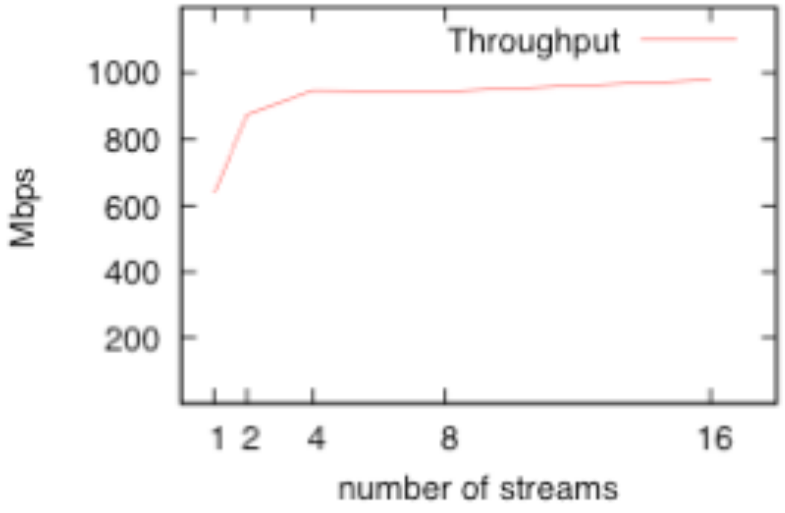


End-to-end Problem

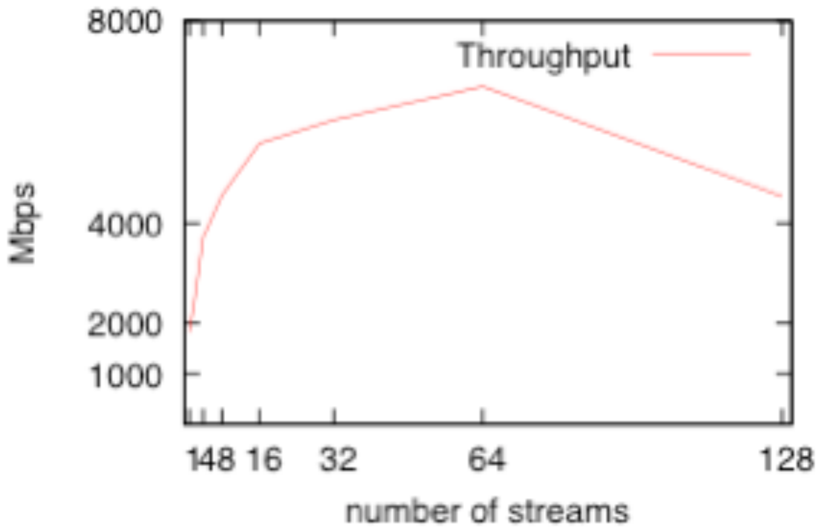


Tnetwork -> Network Throughput
TSmem->network -> Memory-to-network Throughput on source
TSdisk->mem -> Disk-to-memory Throughput on source
TDnetwork->mem -> Network-to-memory Throughput on Destination
TDmem->disk -> Memory-to-disk Throughput on destination

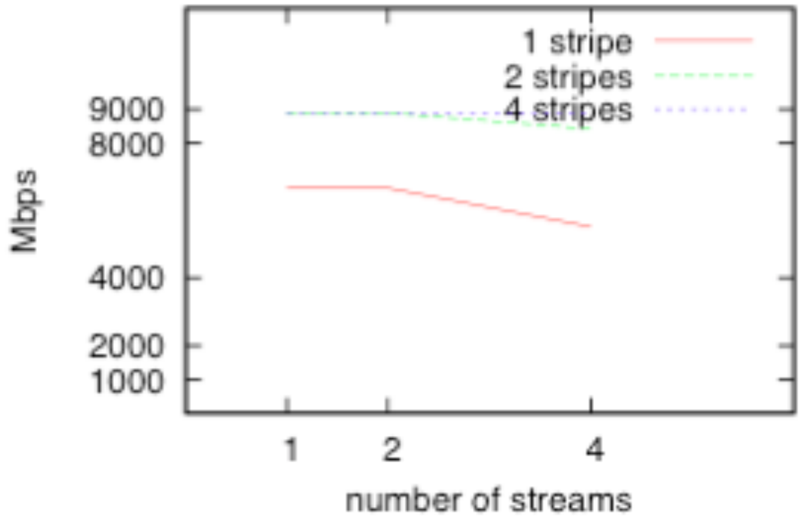
protocol tuning



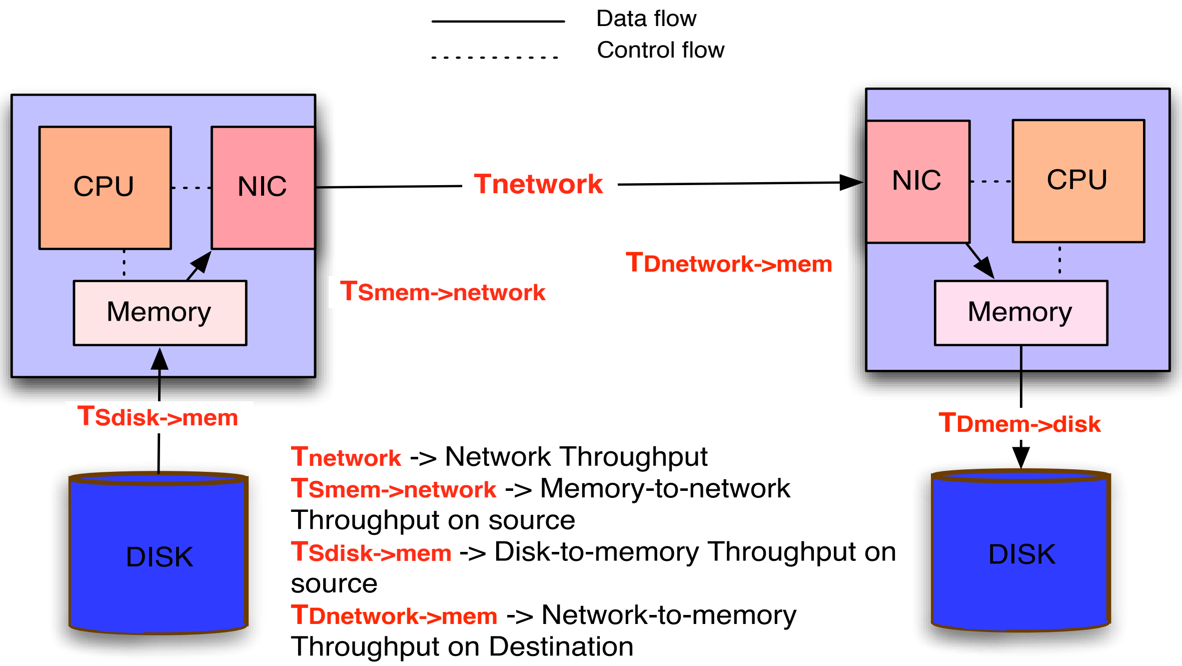
disk I/O optimization



CPU optimization

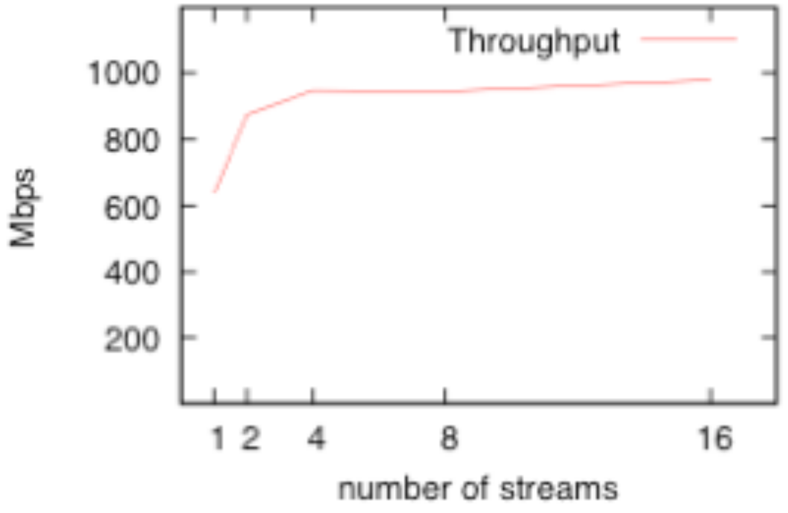


End-to-end Problem

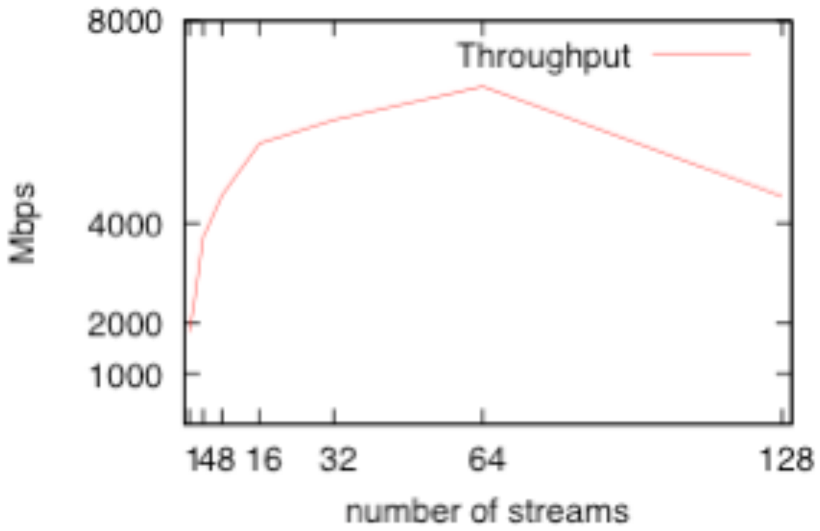


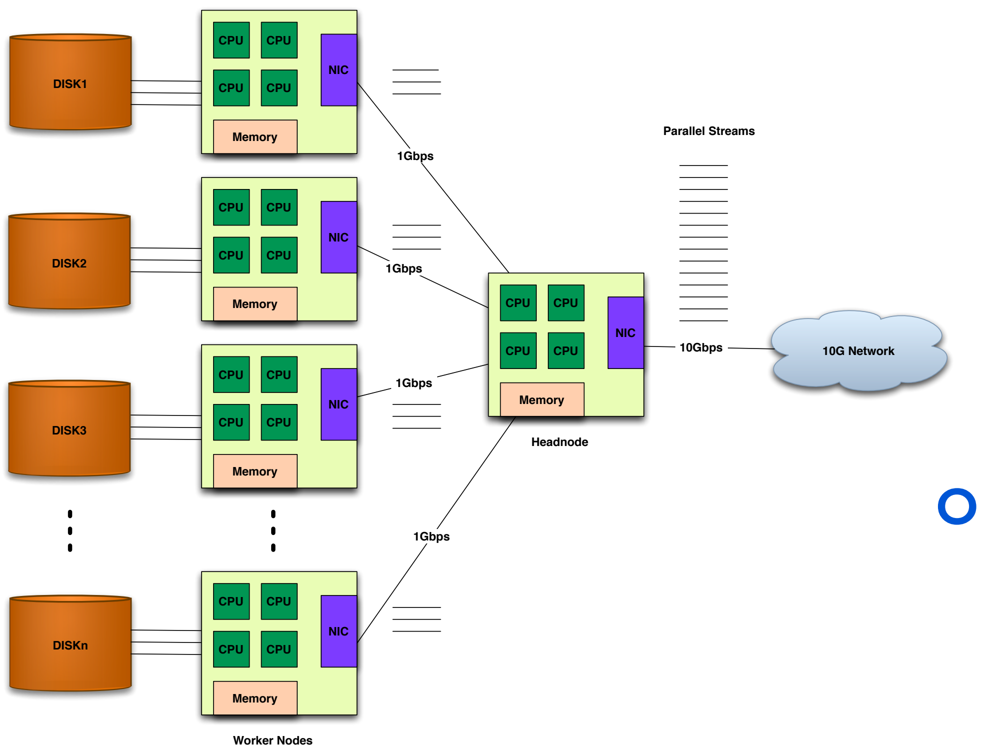
Tnetwork -> Network Throughput
TSmem->network -> Memory-to-network Throughput on source
TSdisk->mem -> Disk-to-memory Throughput on source
TDnetwork->mem -> Network-to-memory Throughput on Destination
TDmem->disk -> Memory-to-disk Throughput on destination

protocol tuning



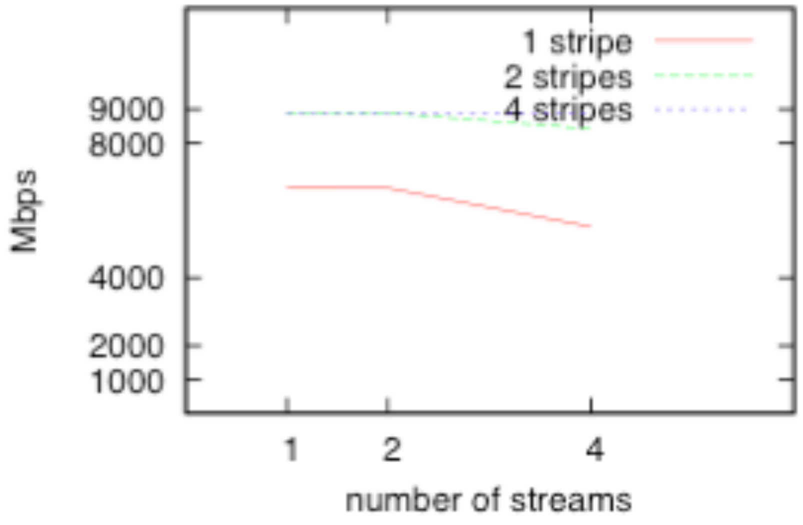
disk I/O optimization



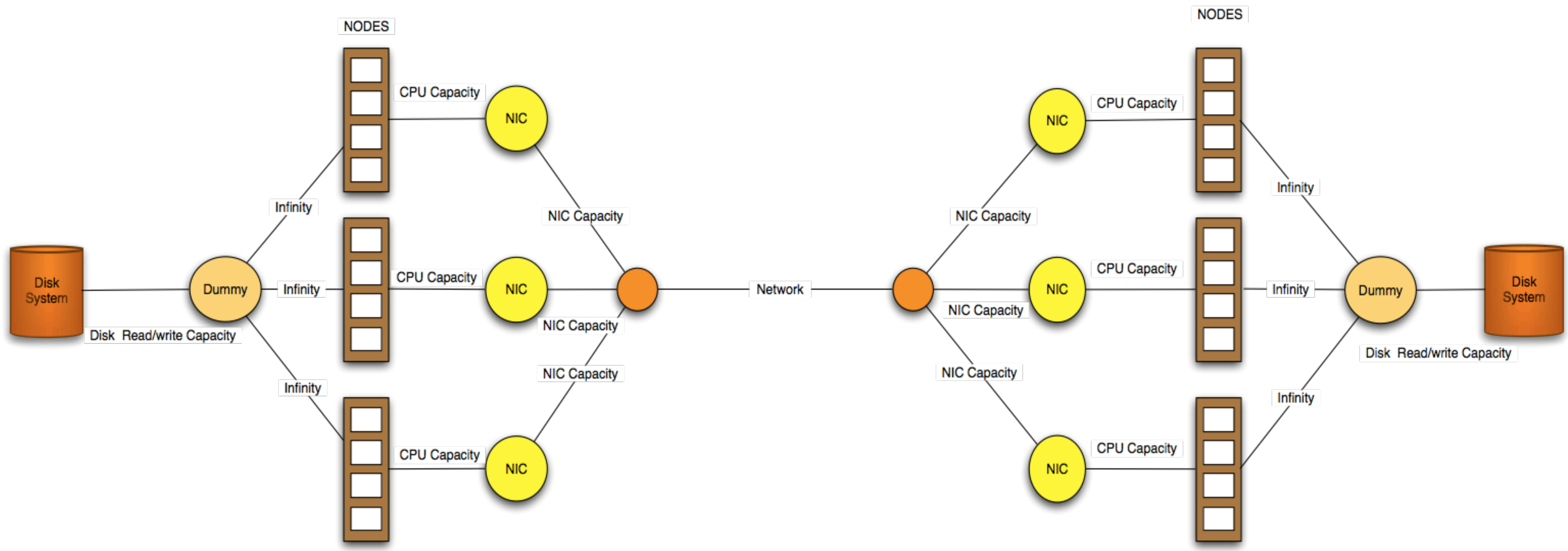


Parameters to be optimized:
- # of streams
- # of disk stripes
- # of CPUs/nodes

CPU optimization



End-to-end Optimization



- CPU nodes are considered as nodes of a maximum flow problem
- Memory-to-memory transfers are simulated with dummy source and sink nodes
- The capacities of disk and network is found by applying parallel stream model by taking into consideration of resource capacities (NIC & CPU)

Challenging Problem

Optimize:

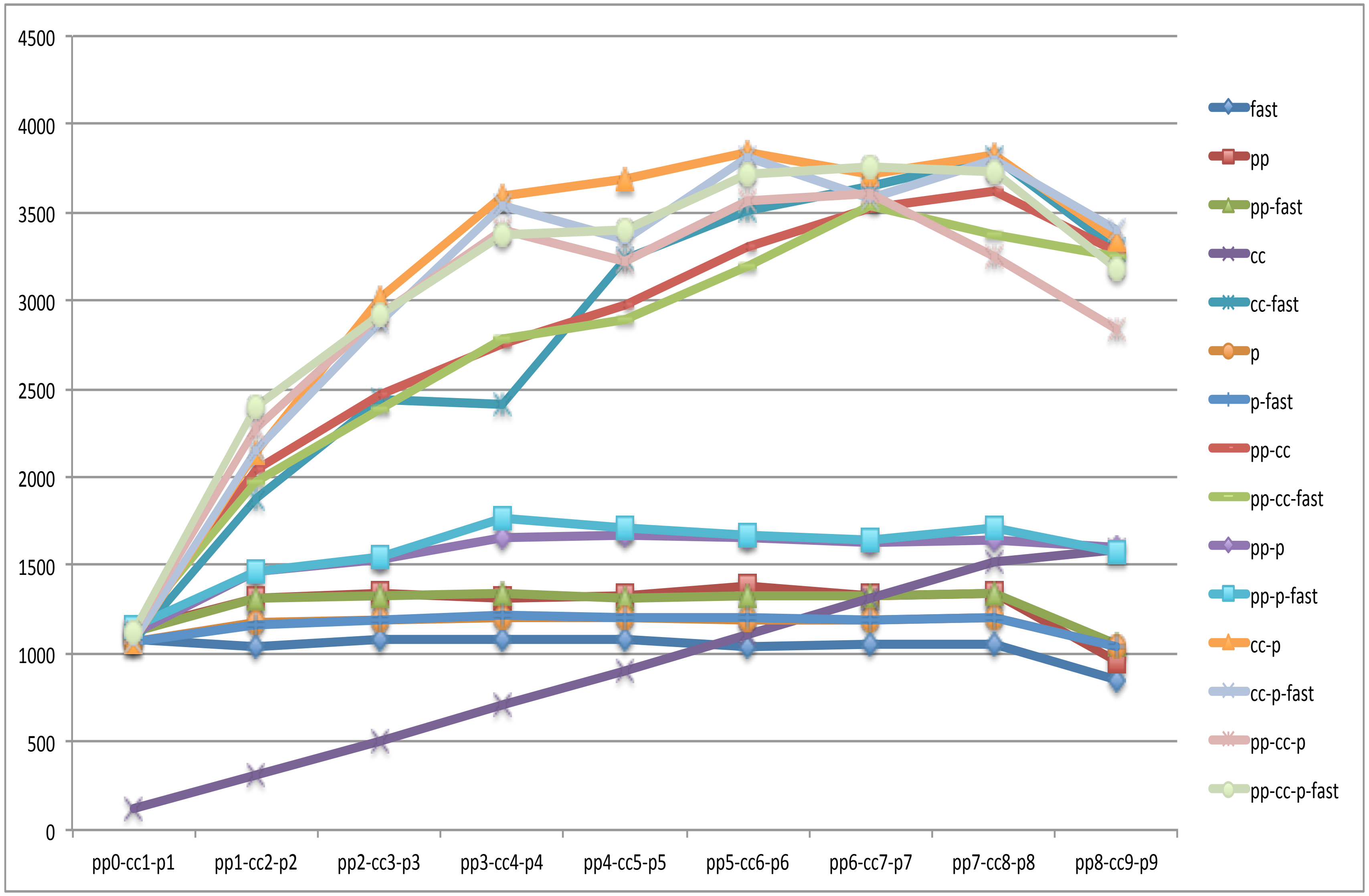
- concurrency
- parallelism
- pipelining
- conn. caching
- buffer size
- block size
- disk striping
- threading
- ....

Challenging Problem

Optimize:

- concurrency
- parallelism
- pipelining
- conn. caching
- buffer size
- block size
- disk striping
- threading
- ....

(1)



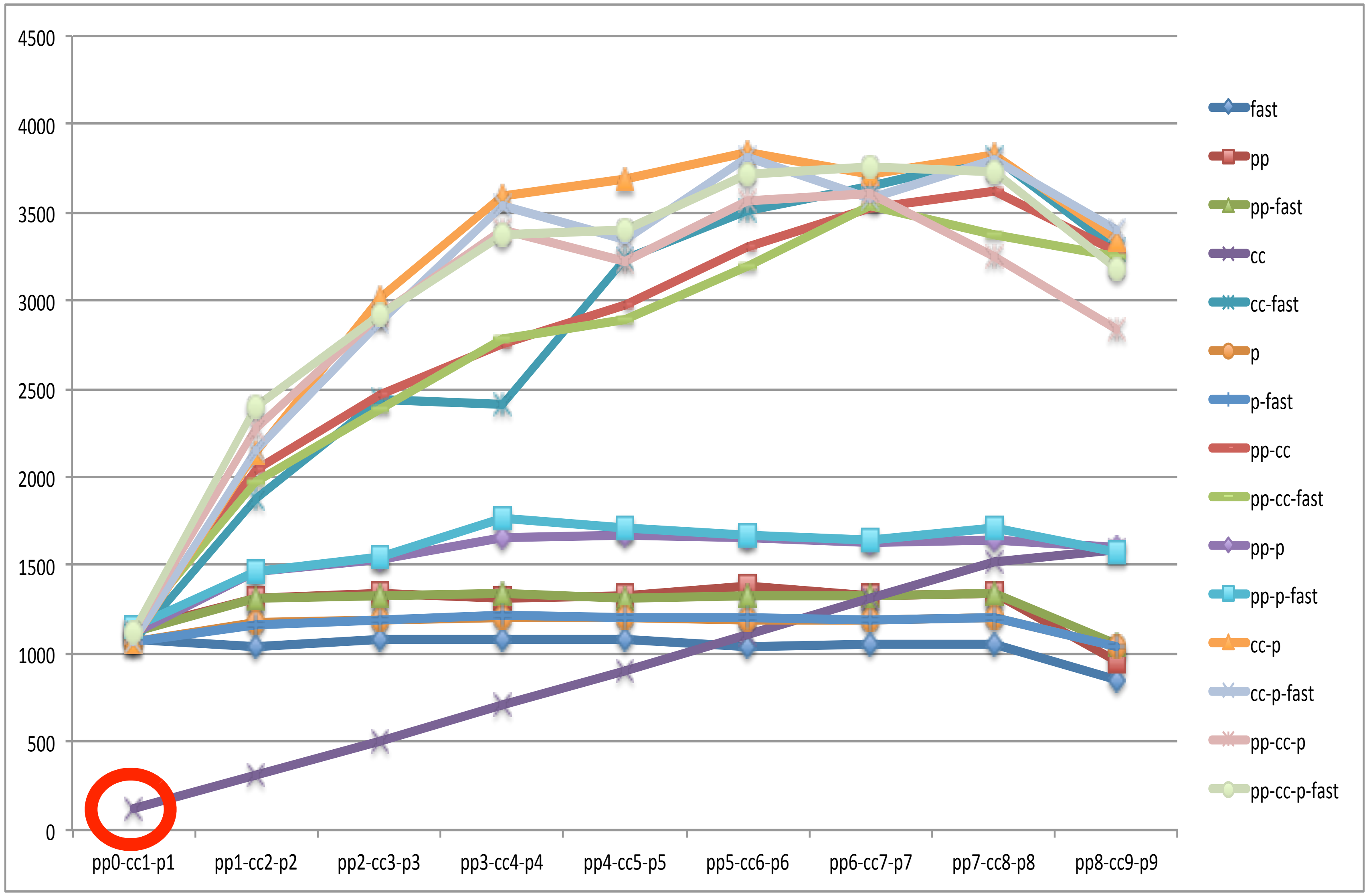
512 x 8 MB files

Challenging Problem

Optimize:

- concurrency
- parallelism
- pipelining
- conn. caching
- buffer size
- block size
- disk striping
- threading
- ....

(1)



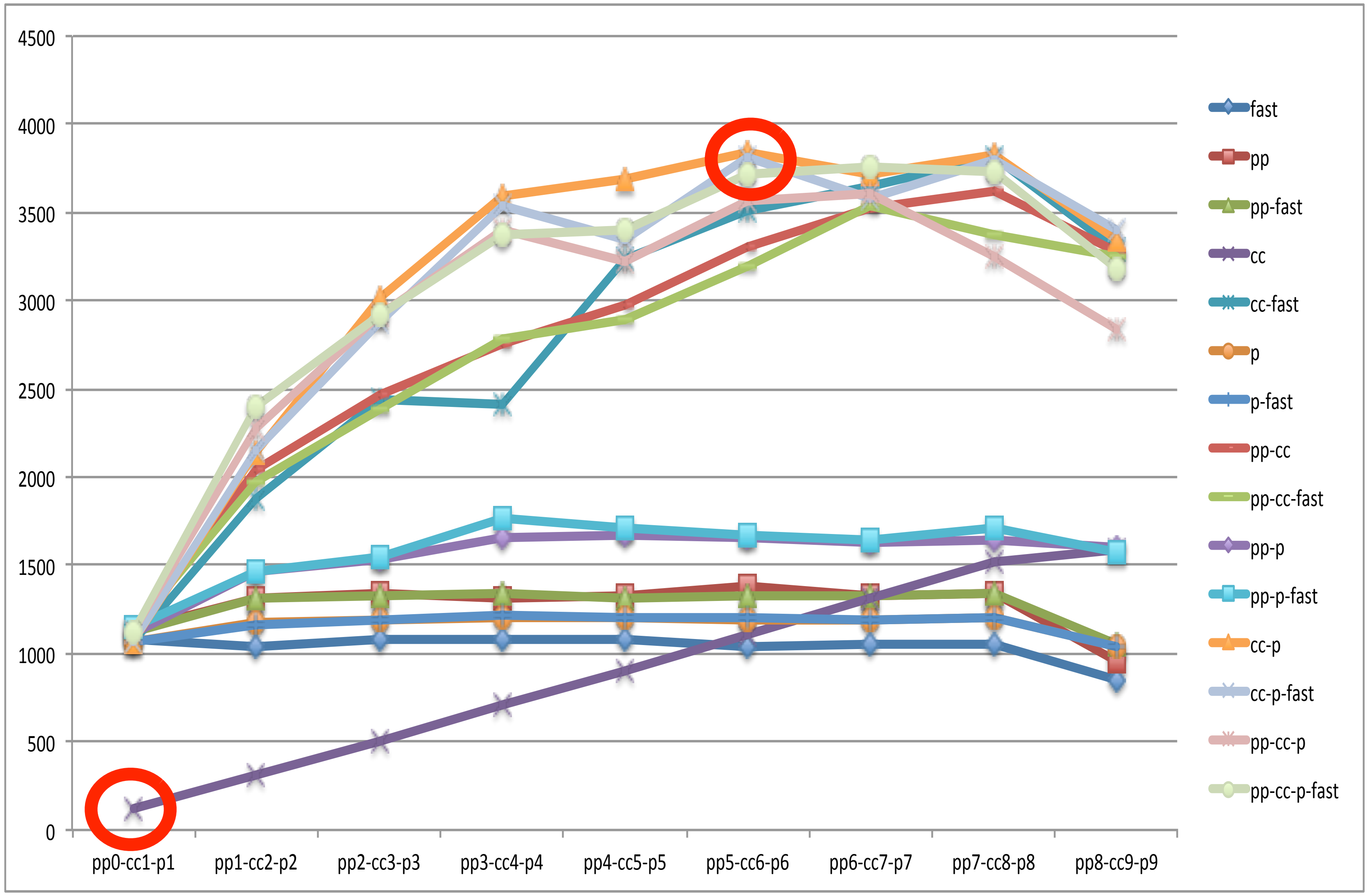
512 x 8 MB files

Challenging Problem

Optimize:

- concurrency
- parallelism
- pipelining
- conn. caching
- buffer size
- block size
- disk striping
- threading
- ....

(1)



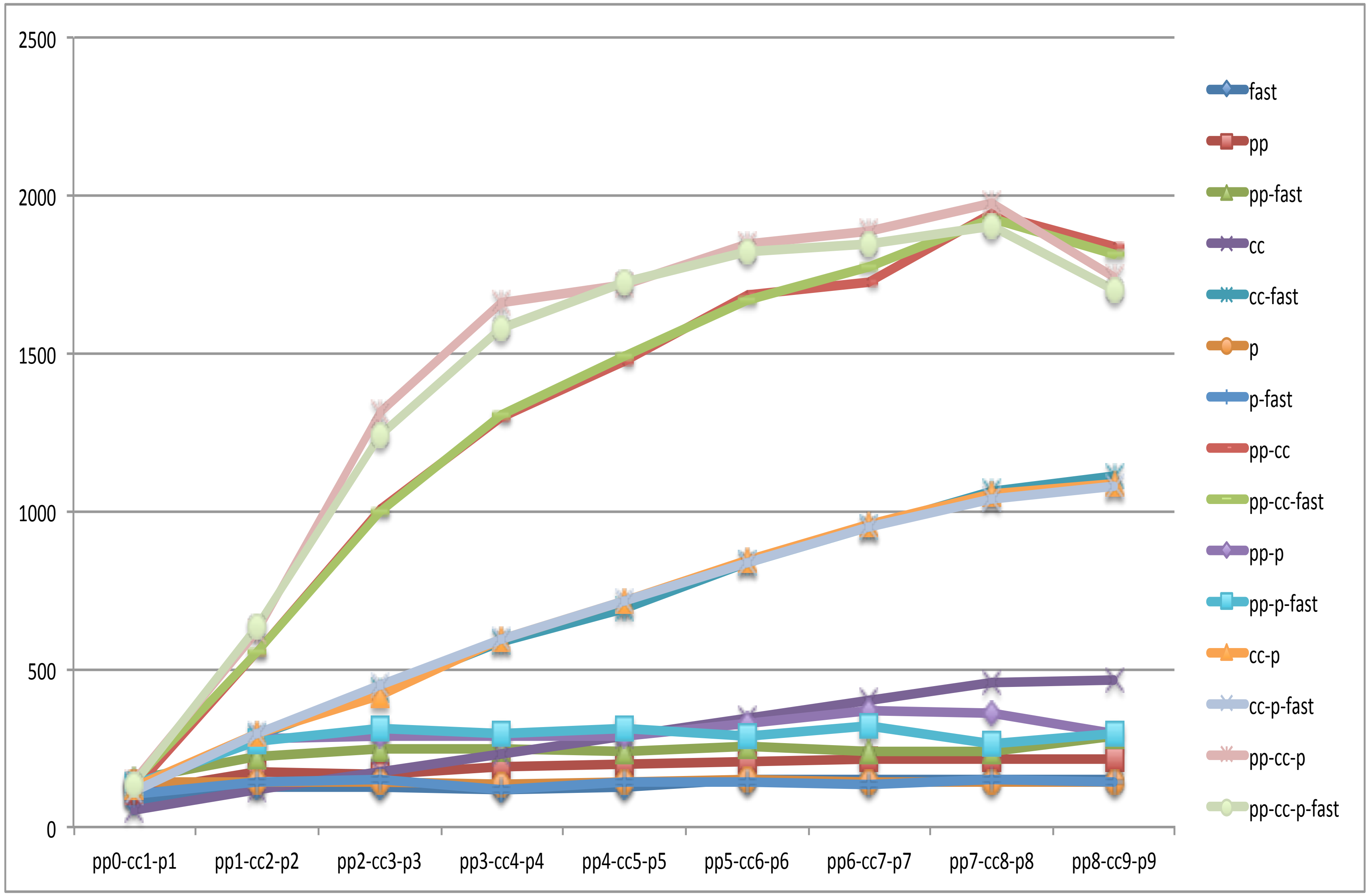
512 x 8 MB files

Challenging Problem

Optimize:

- concurrency
- parallelism
- pipelining
- conn. caching
- buffer size
- block size
- disk striping
- threading
- ....

(2)



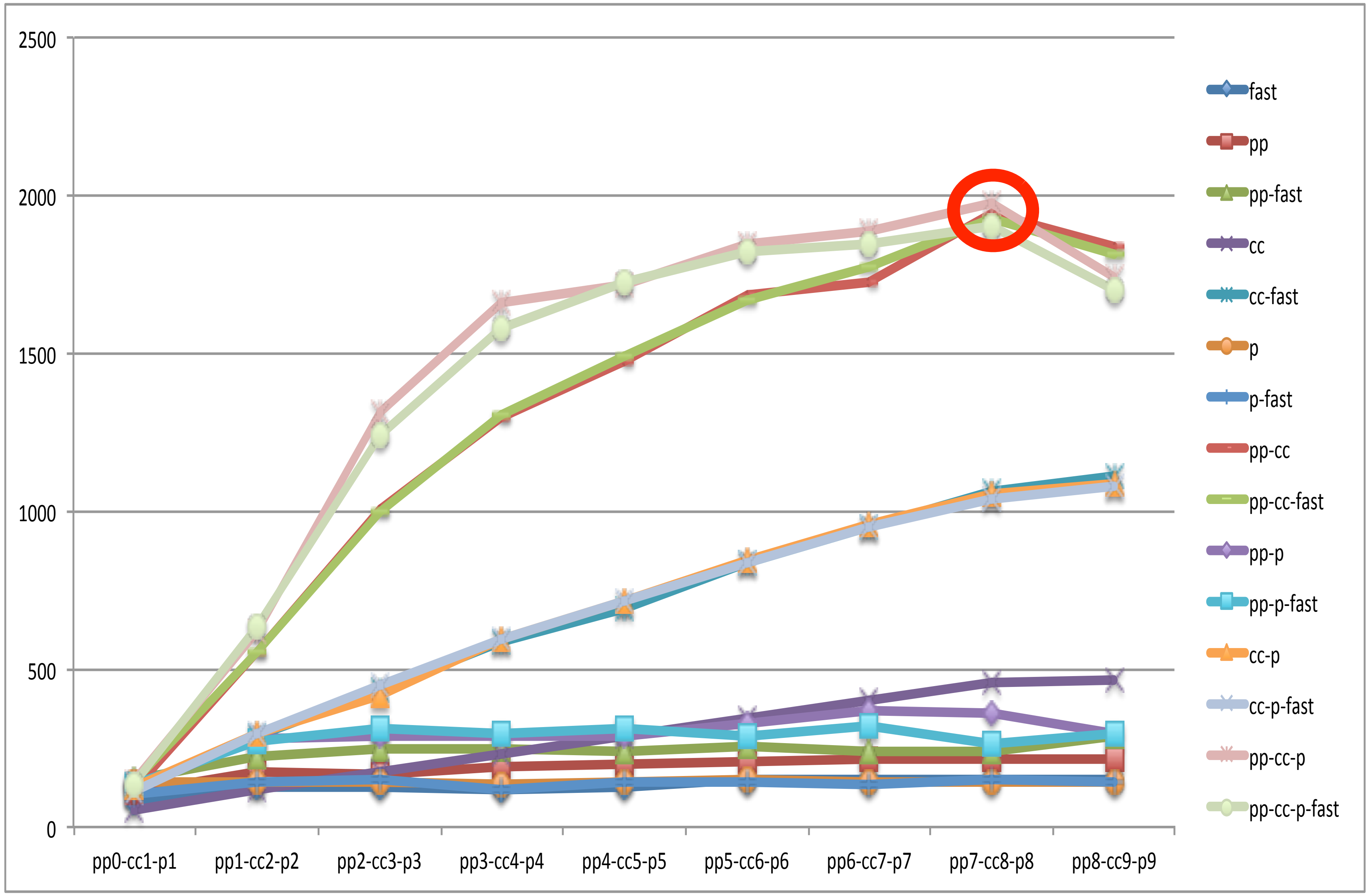
512 x 1 MB files

Challenging Problem

Optimize:

- concurrency
- parallelism
- pipelining
- conn. caching
- buffer size
- block size
- disk striping
- threading
- ….

(2)



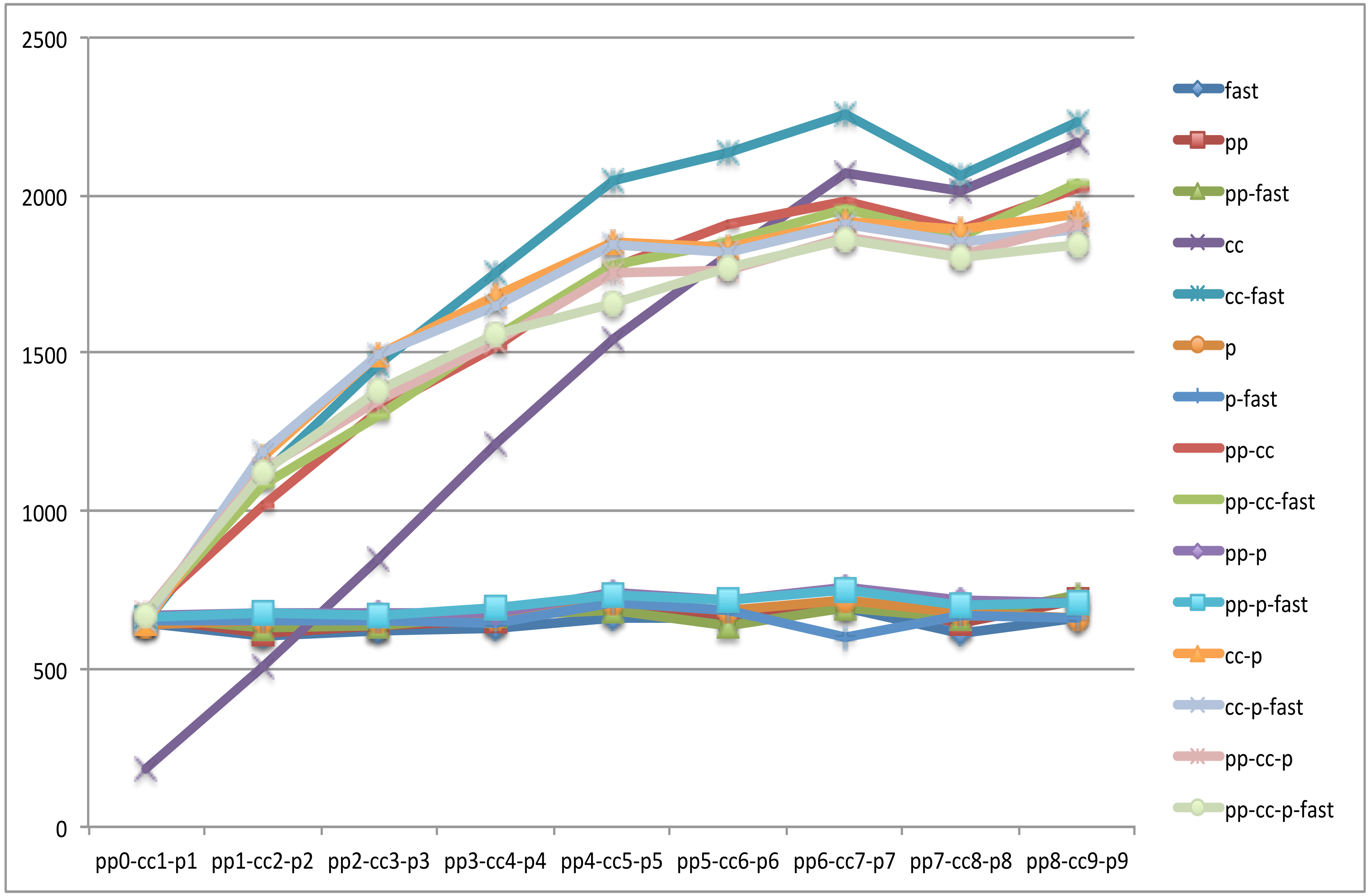
512 x 1 MB files

Challenging Problem

Optimize:

- concurrency
- parallelism
- pipelining
- conn. caching
- buffer size
- block size
- disk striping
- threading
- ….

(3)



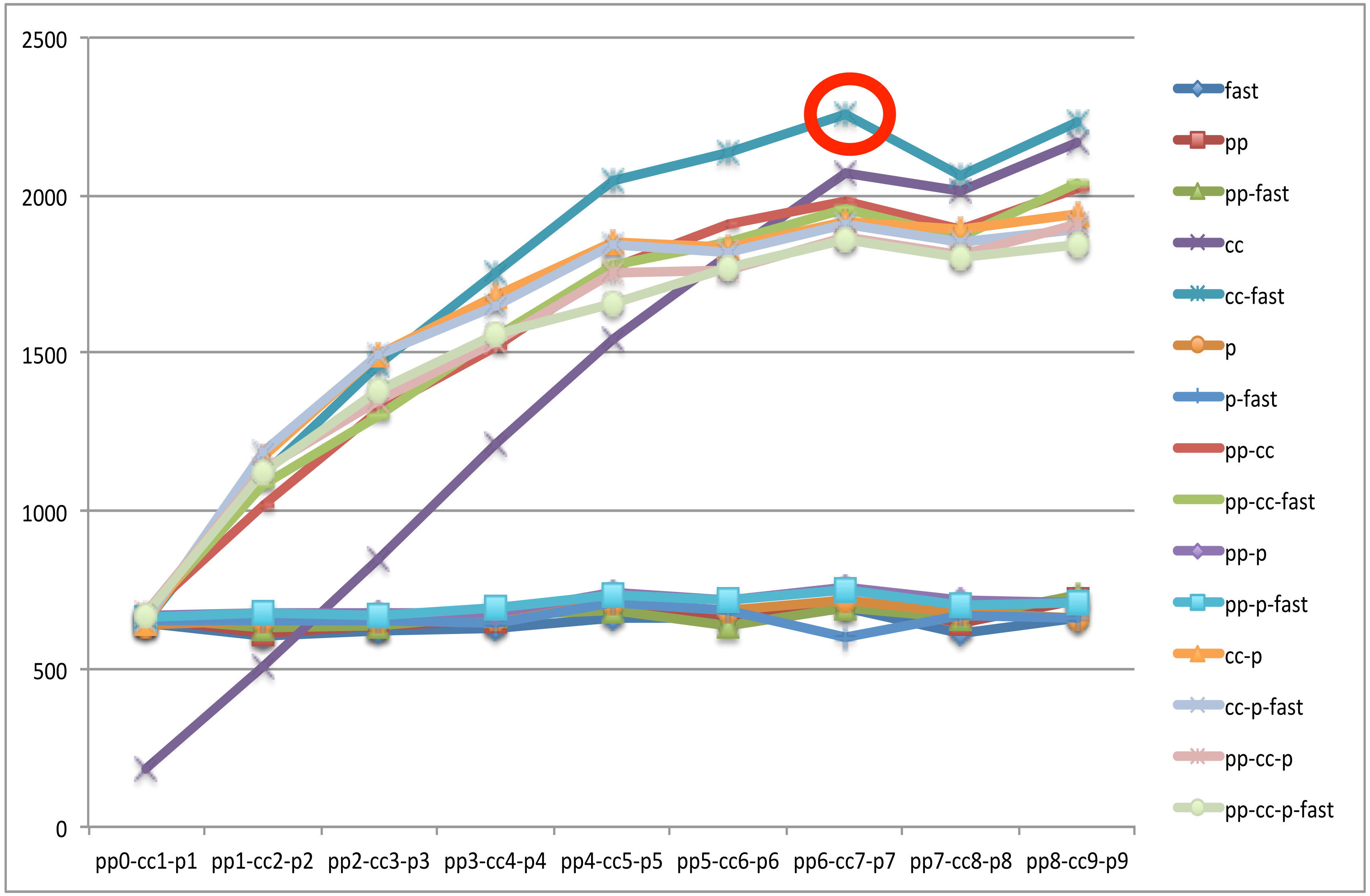
512 x 32 MB files

Challenging Problem

Optimize:

- concurrency
- parallelism
- pipelining
- conn. caching
- buffer size
- block size
- disk striping
- threading
- ....

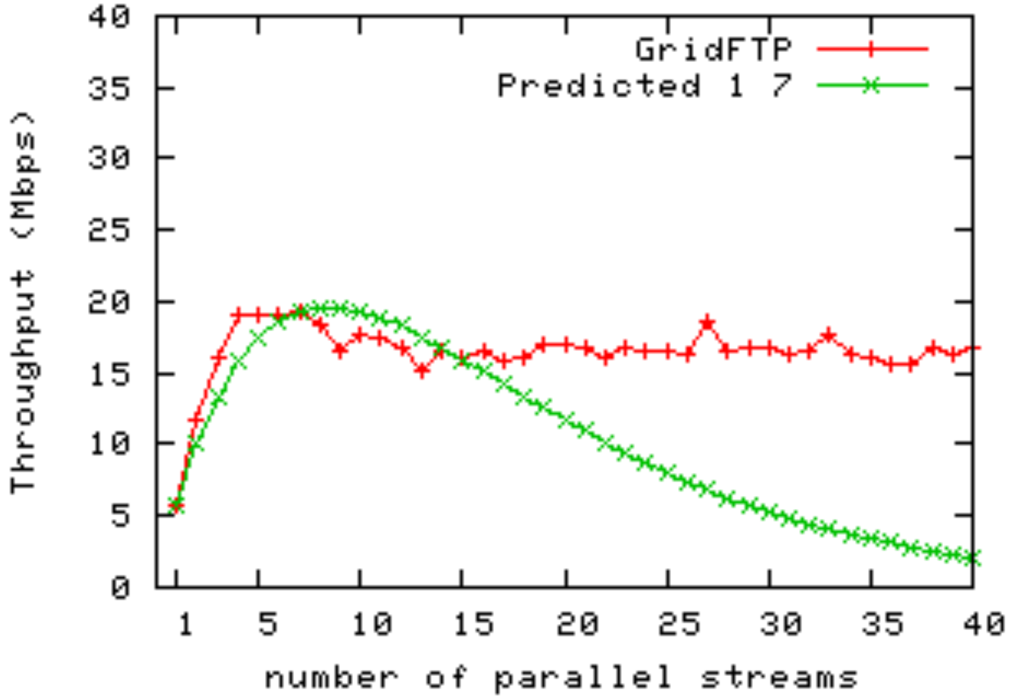
(3)



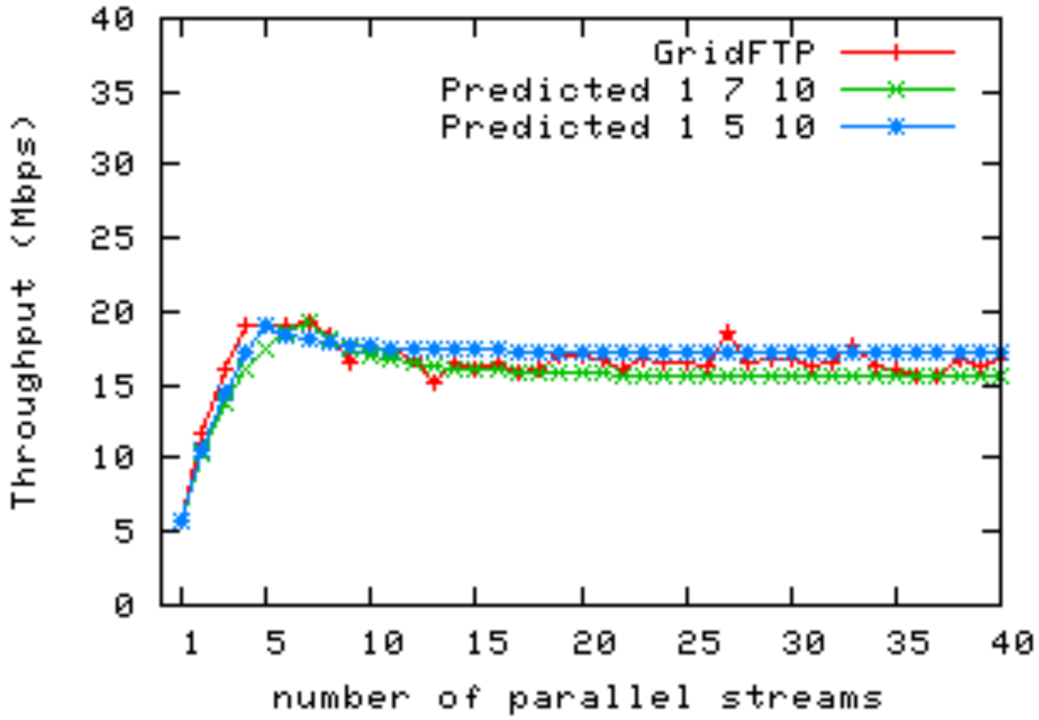
512 x 32 MB files

Kosar et al Models

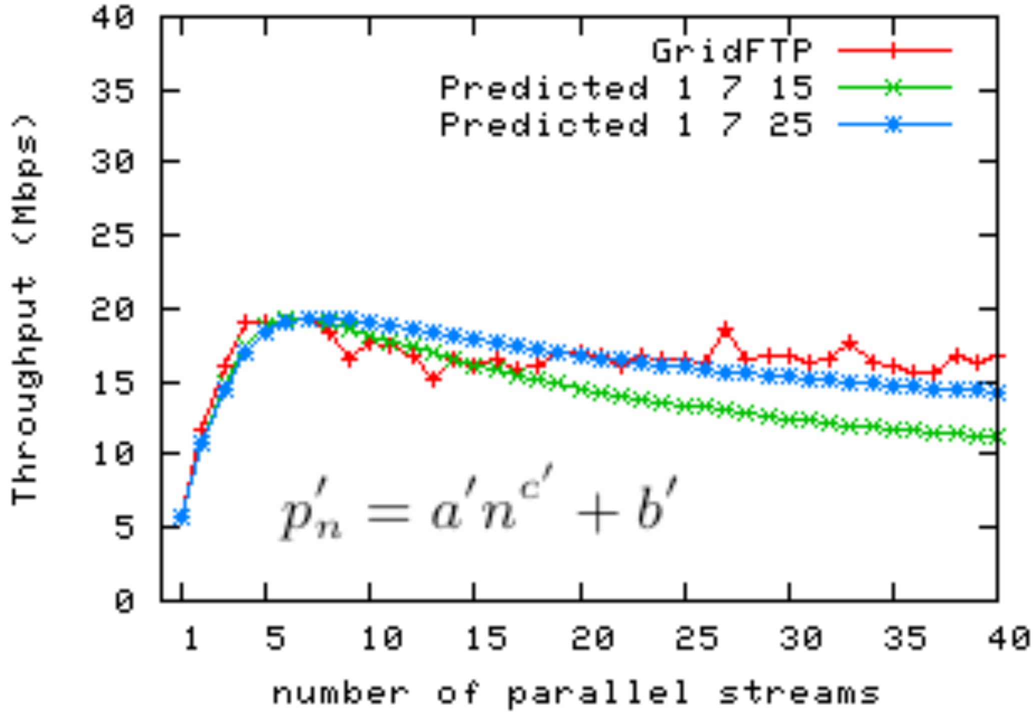
Exponential Packet Loss



Break Function Modeling

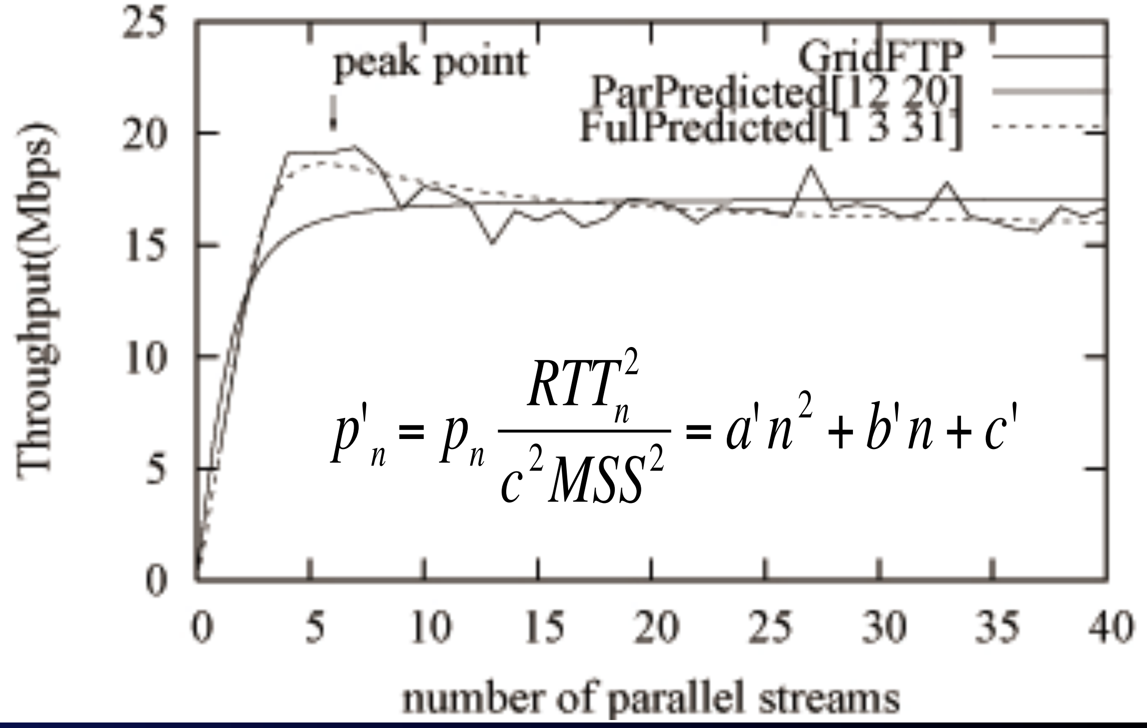


Modeling Based on Newton’s Iteration



$$p'_n = a'n^{c'} + b'$$

Modeling Based on Full Second Order

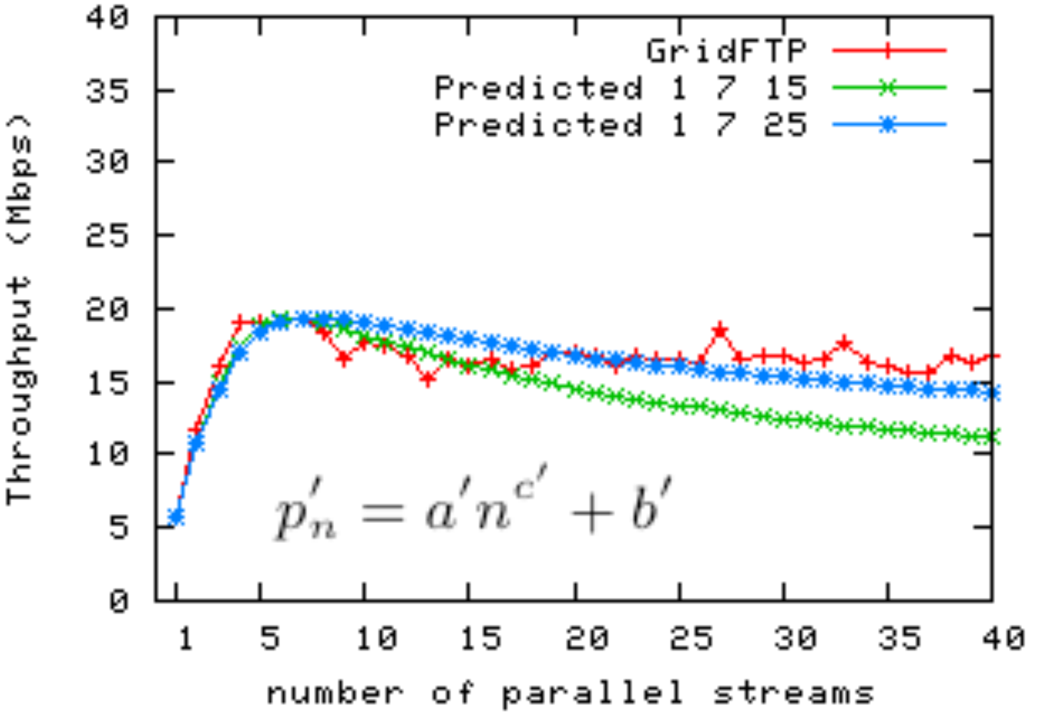


$$p'_n = p_n \frac{RTT_n^2}{c^2 MSS^2} = a'n^2 + b'n + c'$$

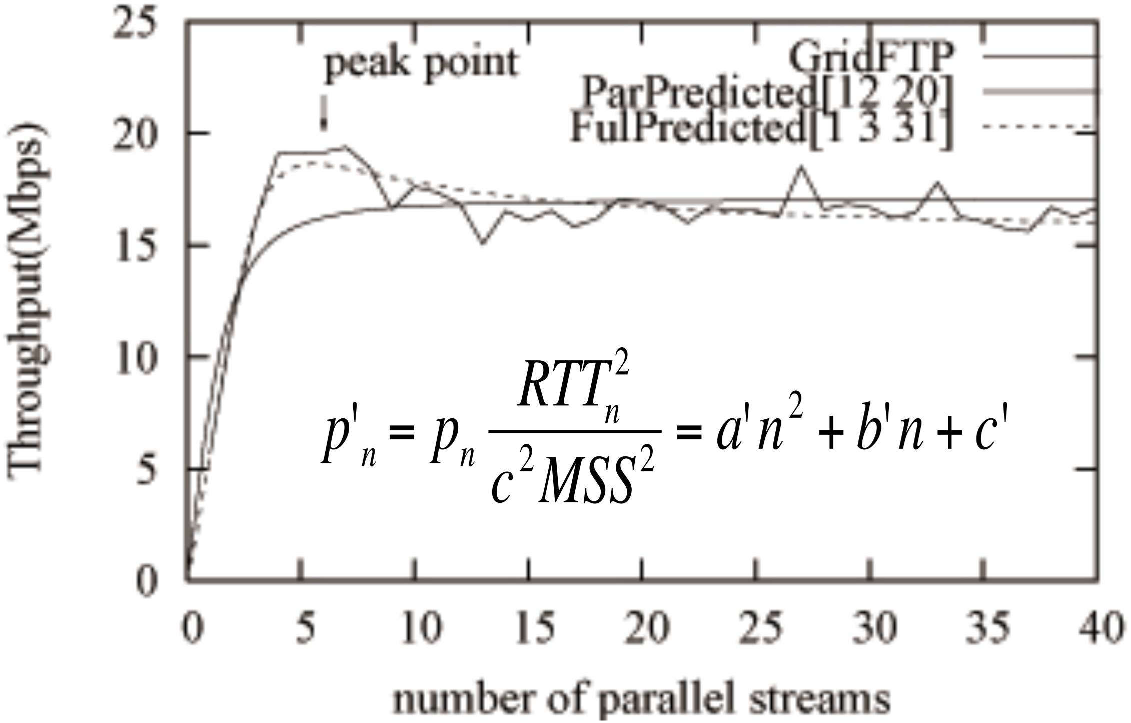
Kosar et al Models

- Details in 2 TPDS 2011 papers
- Implemented in the latest version of Stork (v.2.0.1)
- Provides throughput optimization as well as estimation

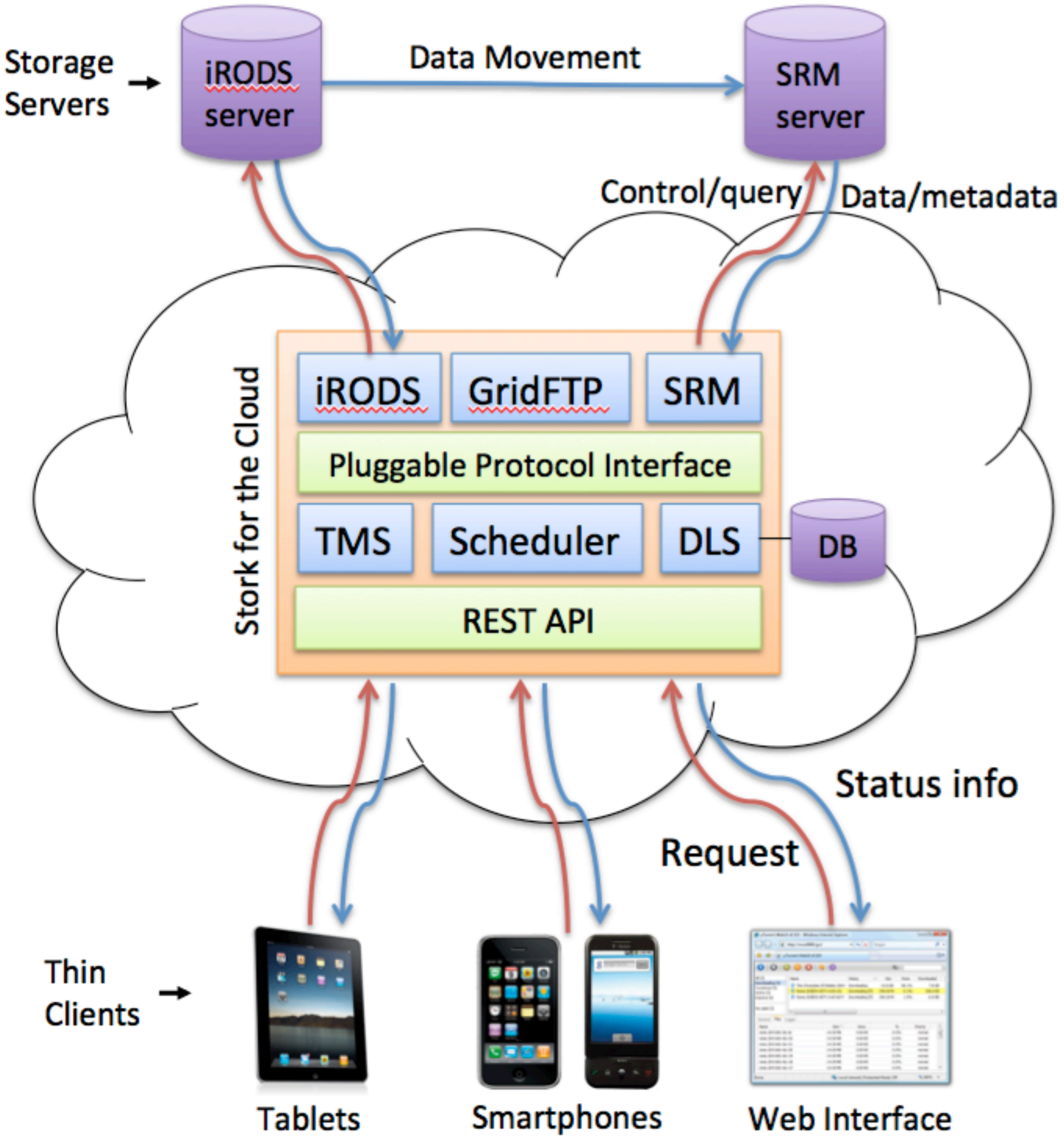
Modeling Based on Newton’s Iteration



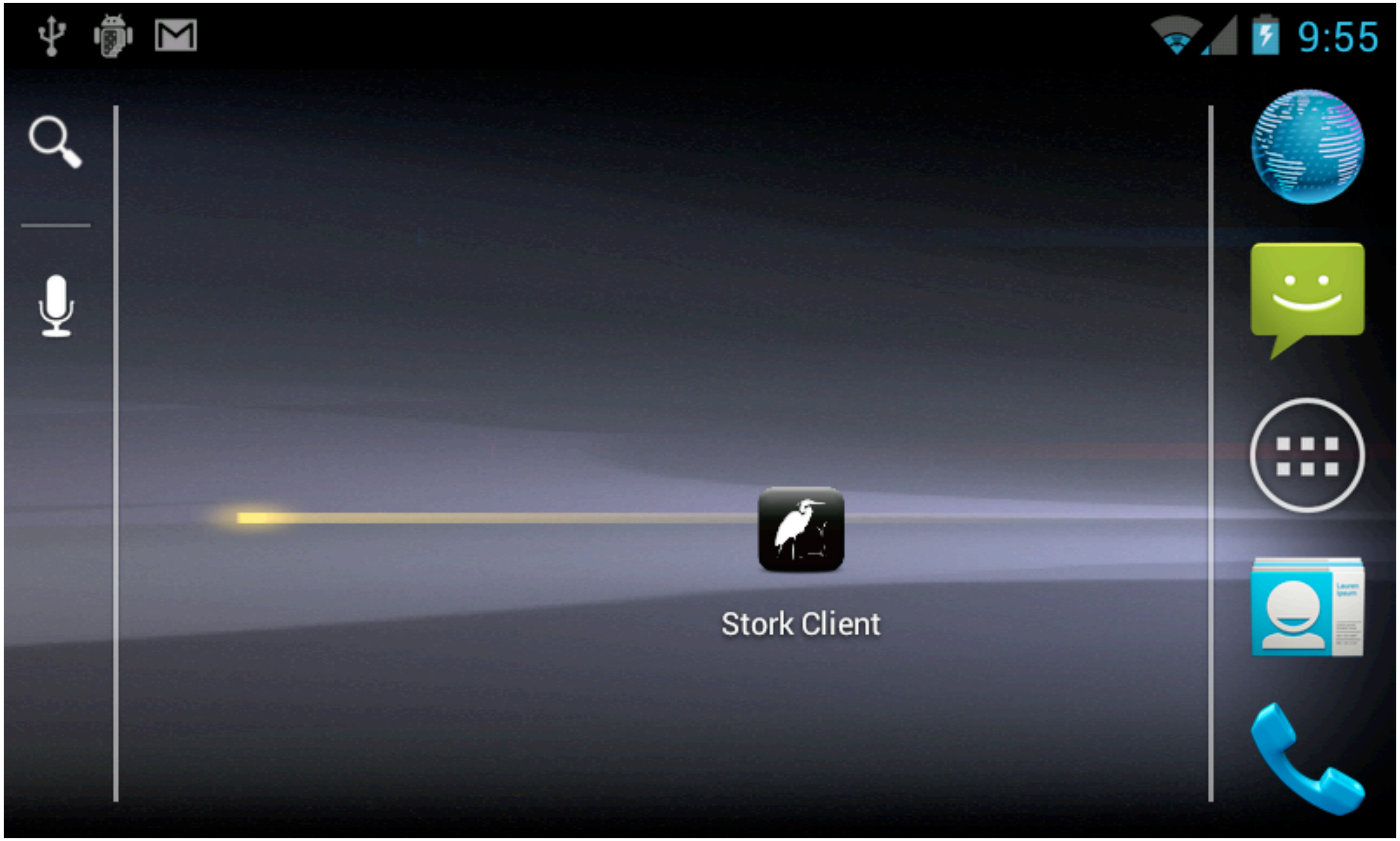
Modeling Based on Full Second Order



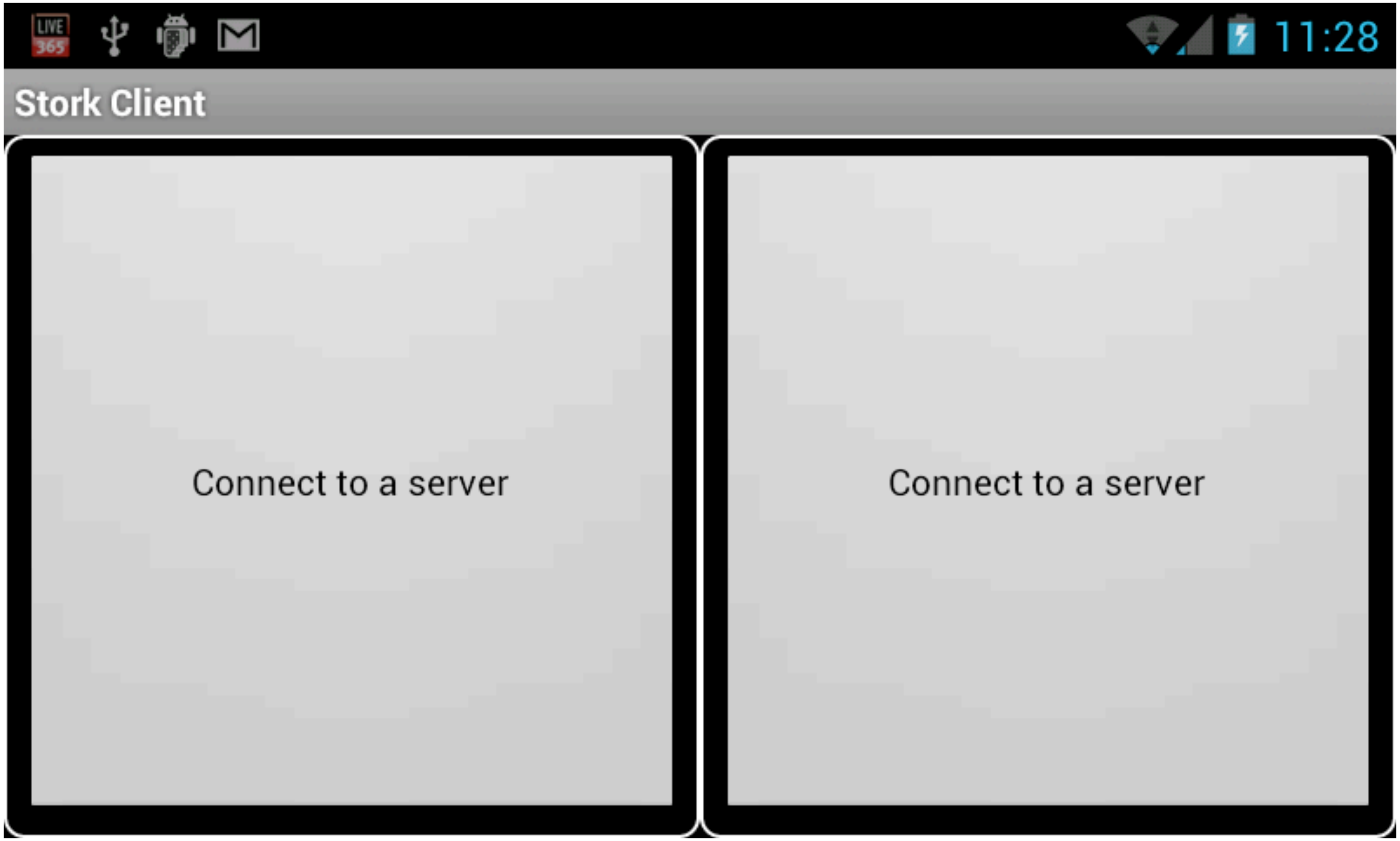
Stork for the Cloud



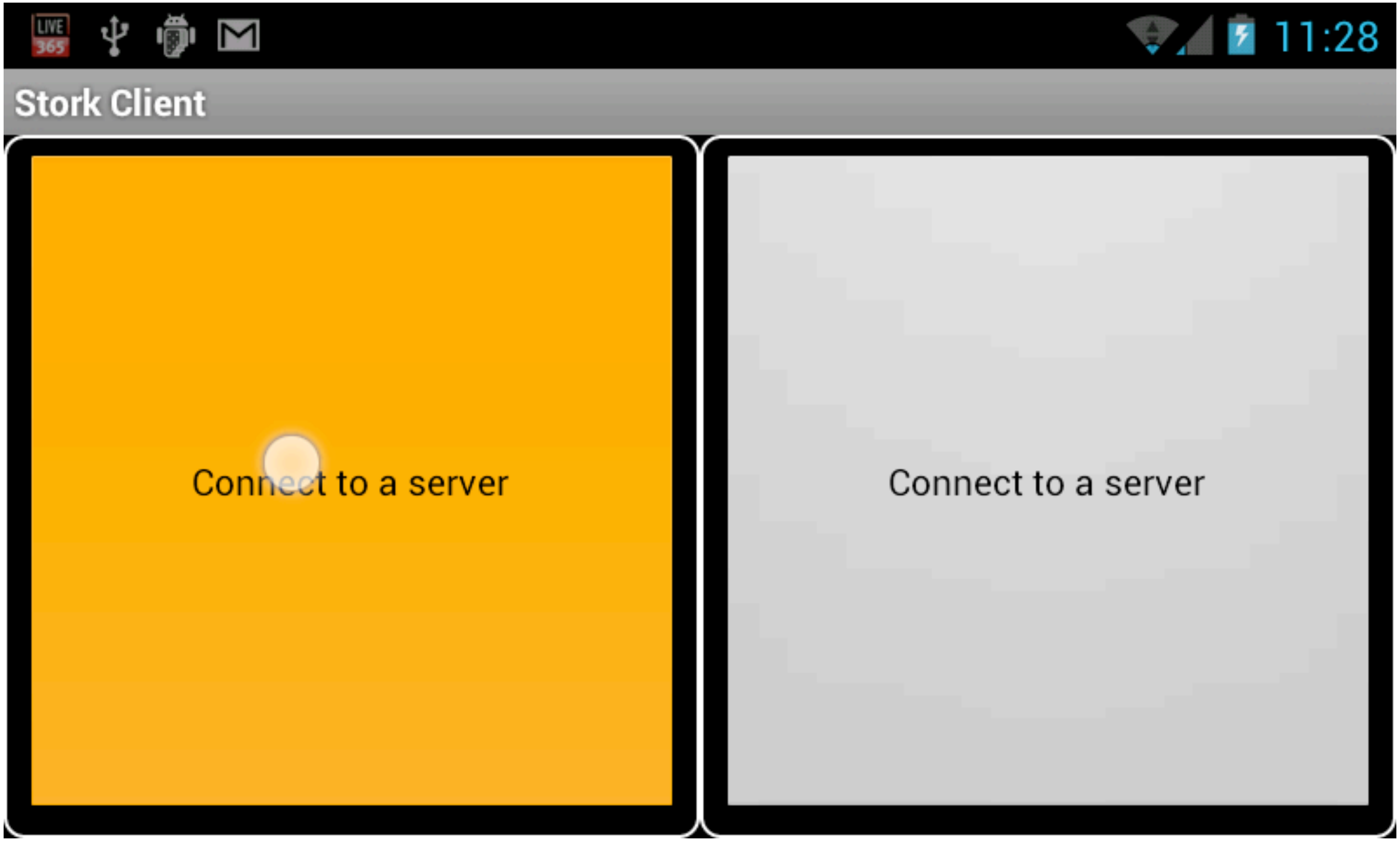
Stork Android Client



Stork Android Client



Stork Android Client

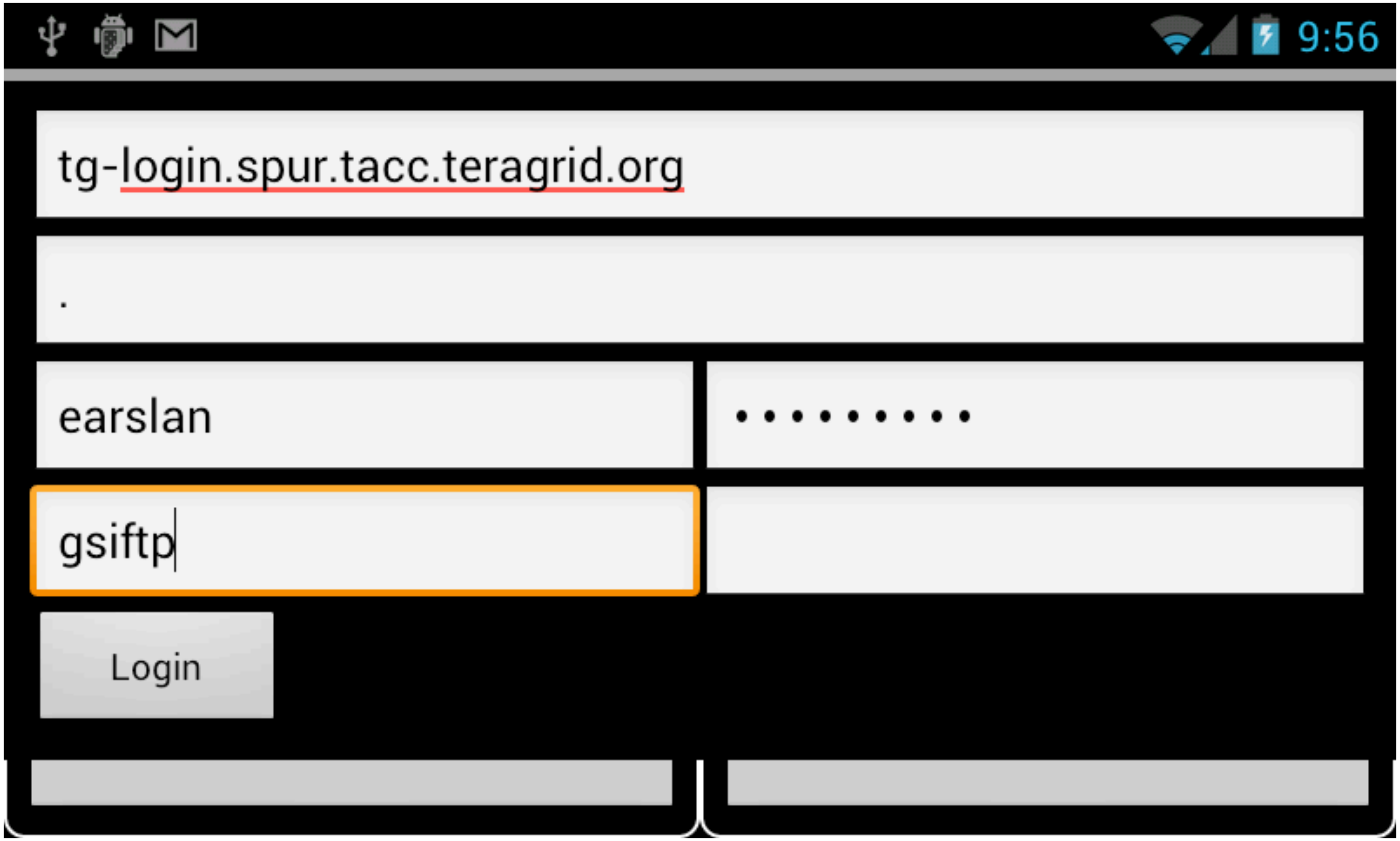


Stork Client

Connect to a server

Connect to a server

Stork Android Client



9:56

tg-login.spur.tacc.teragrid.org

.

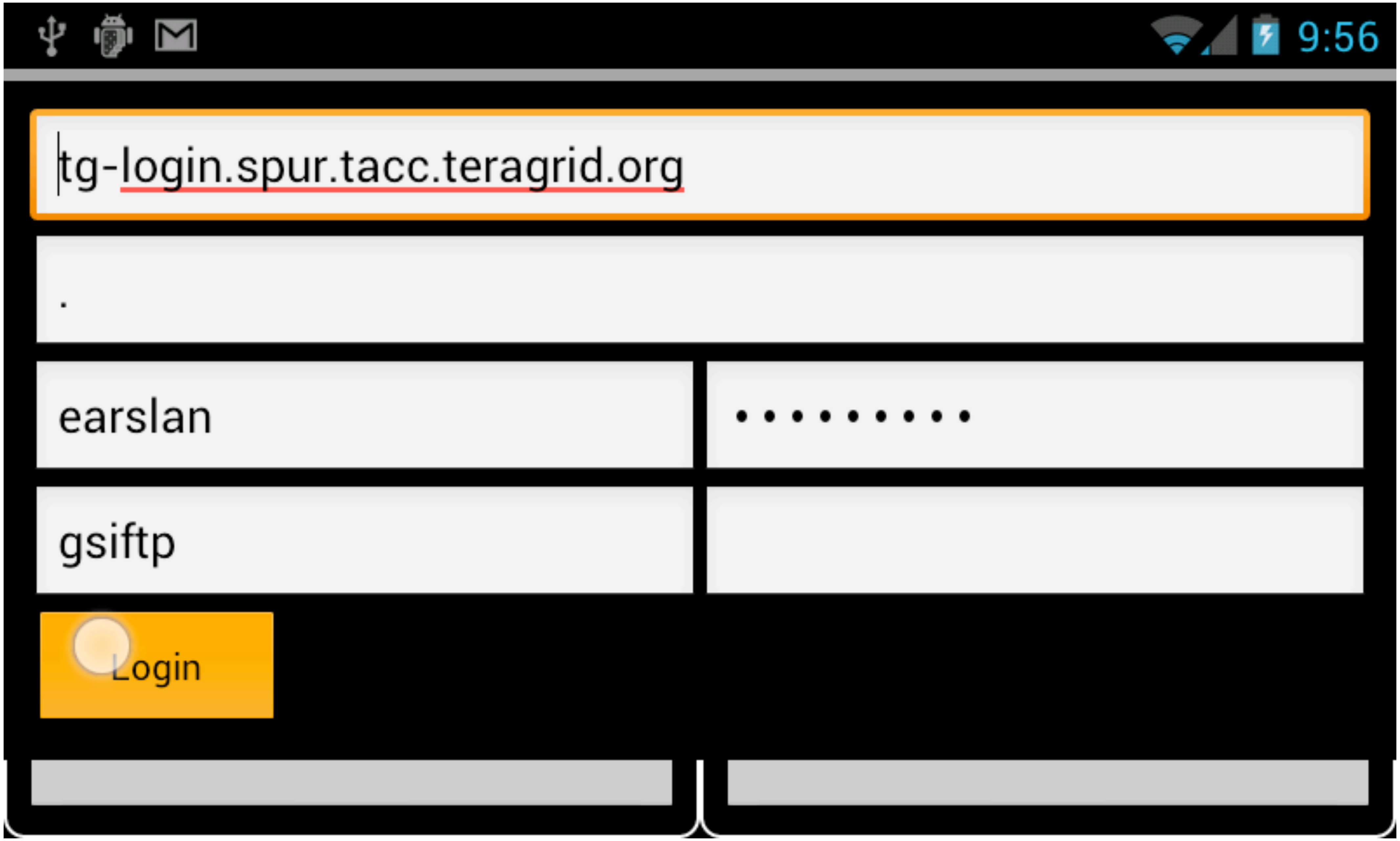
earslan

••••••••••

gsiftp

Login

Stork Android Client



9:56

tg-login.spur.tacc.teragrid.org

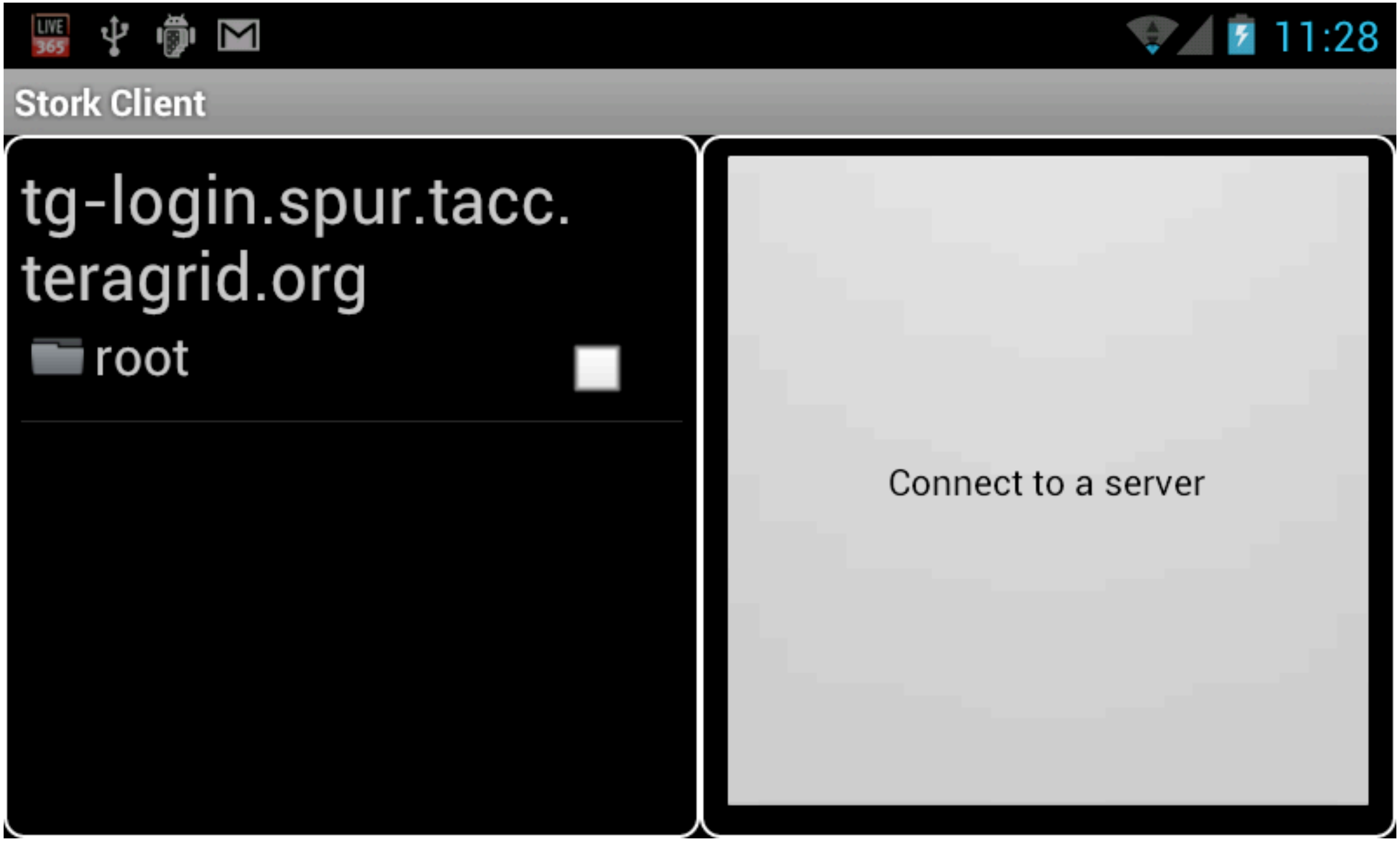
.

earslan

••••••••••

gsiftp

Stork Android Client



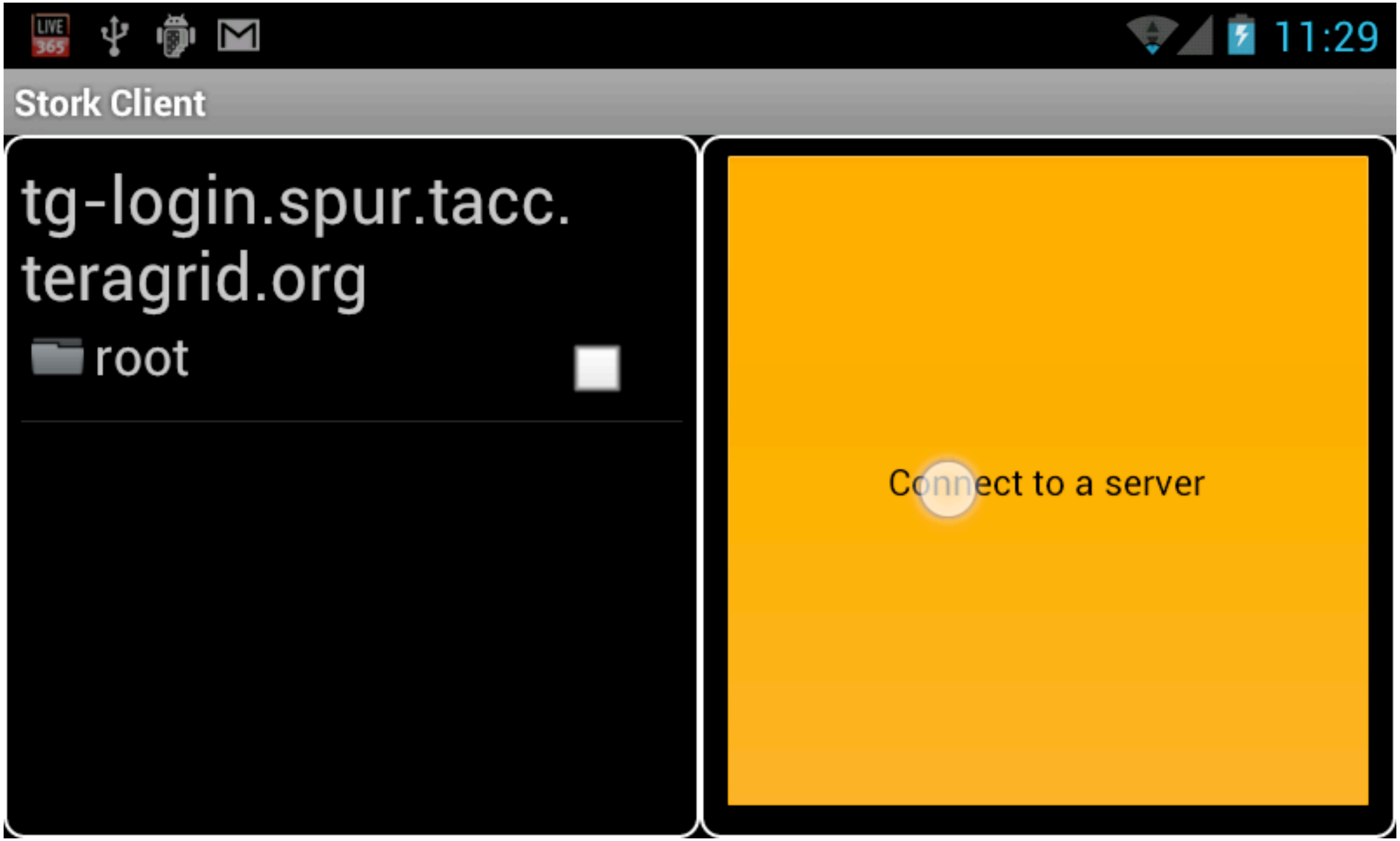
Stork Client

tg-login.spur.tacc.teragrid.org

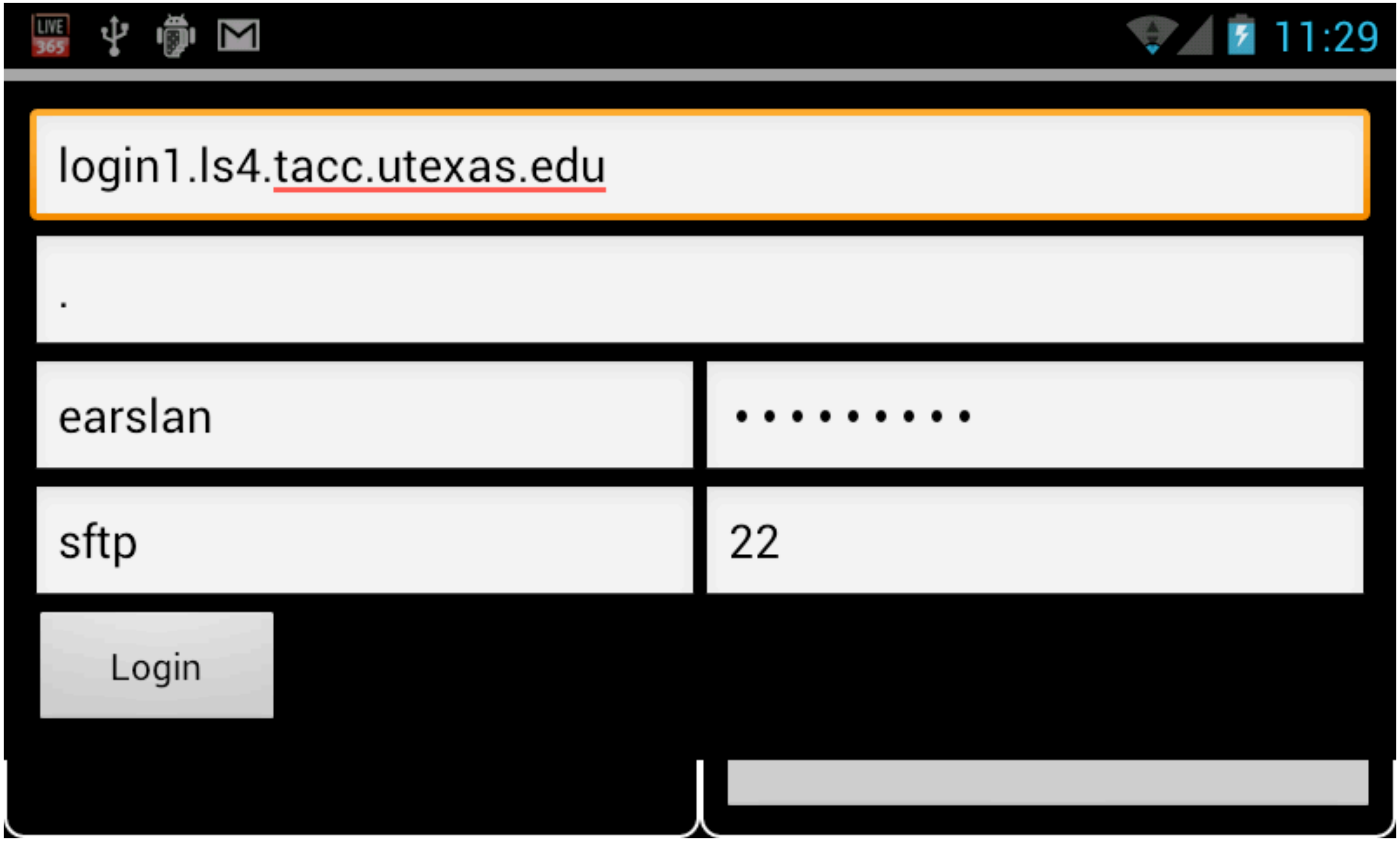
root □

Connect to a server

Stork Android Client



Stork Android Client



login1.ls4.tacc.utexas.edu

.

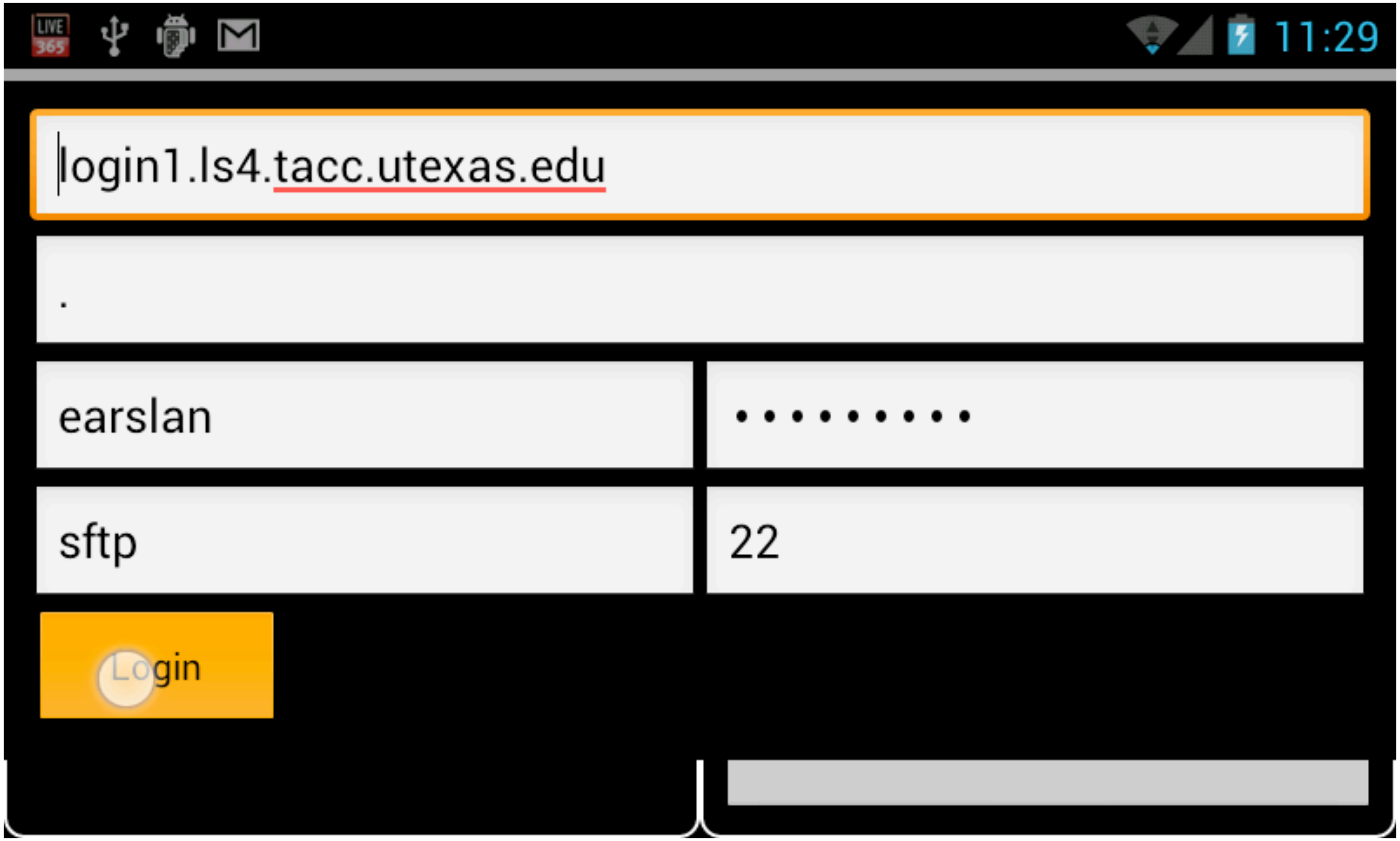
earslan

••••••••••

sftp

22

Stork Android Client



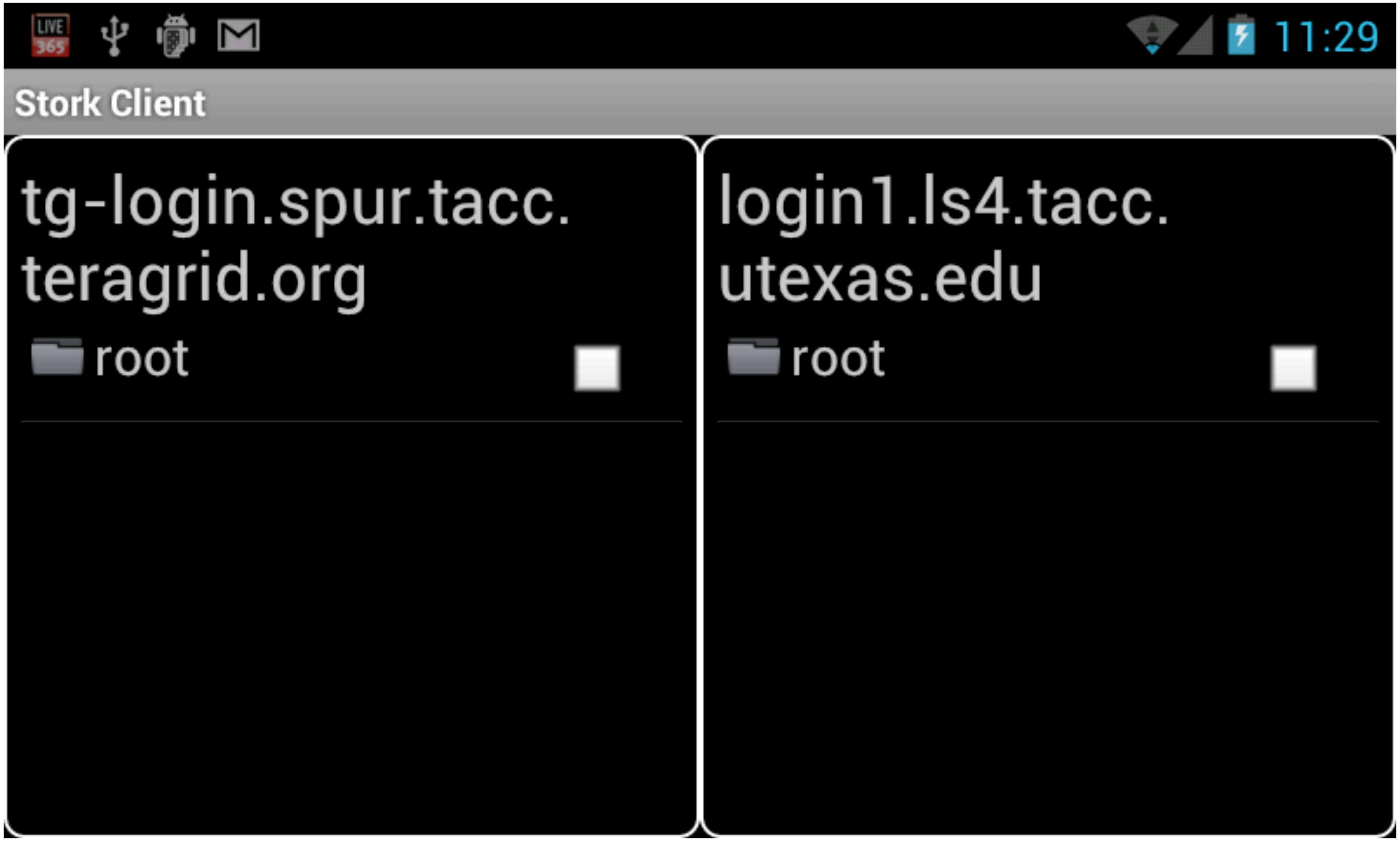
login1.ls4.tacc.utexas.edu

.

earslan •••••••••

sftp 22

Stork Android Client



11:29

Stork Client

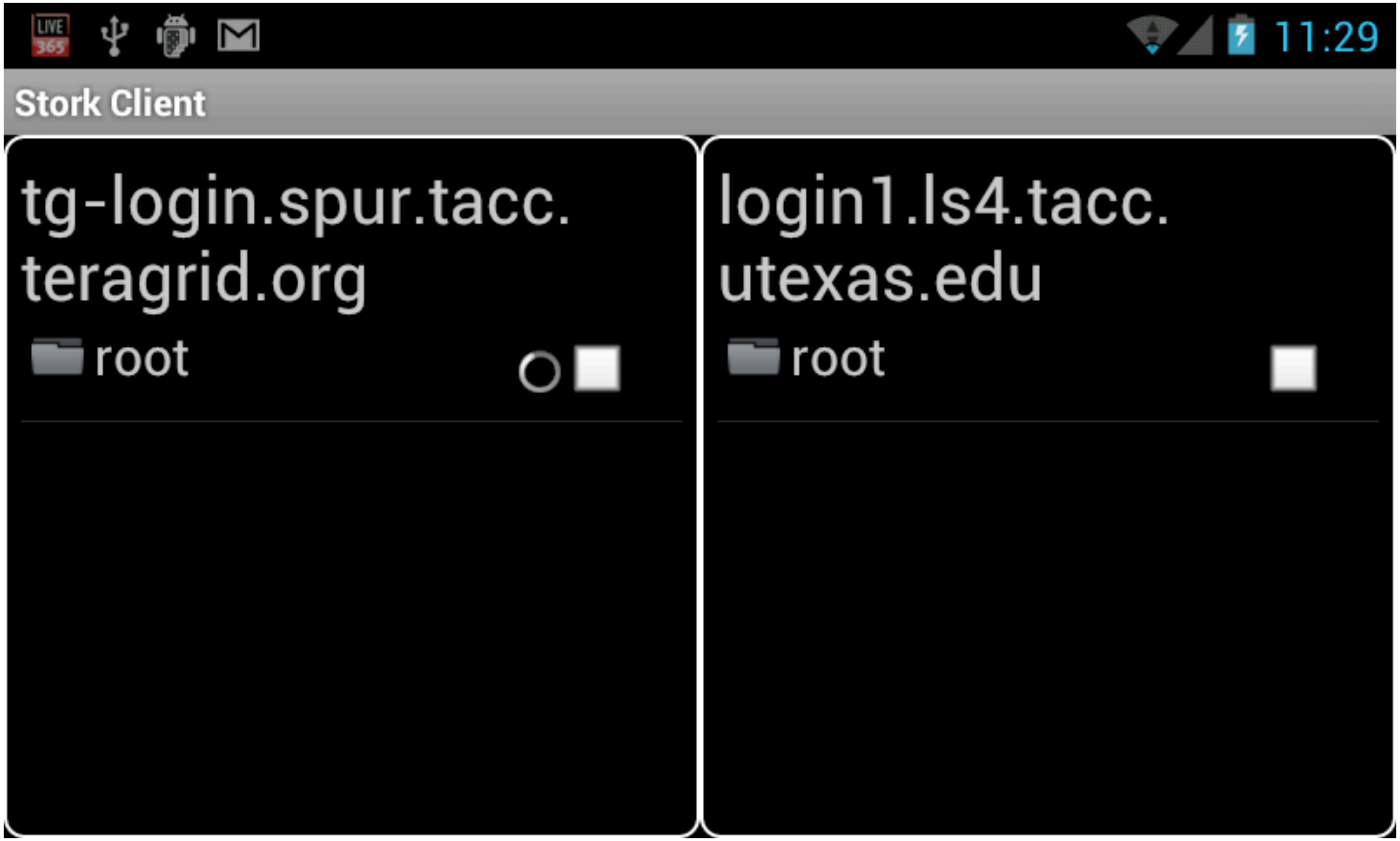
tg-login.spur.tacc.teragrid.org

root □

login1.ls4.tacc.utexas.edu

root □

Stork Android Client



11:29

Stork Client

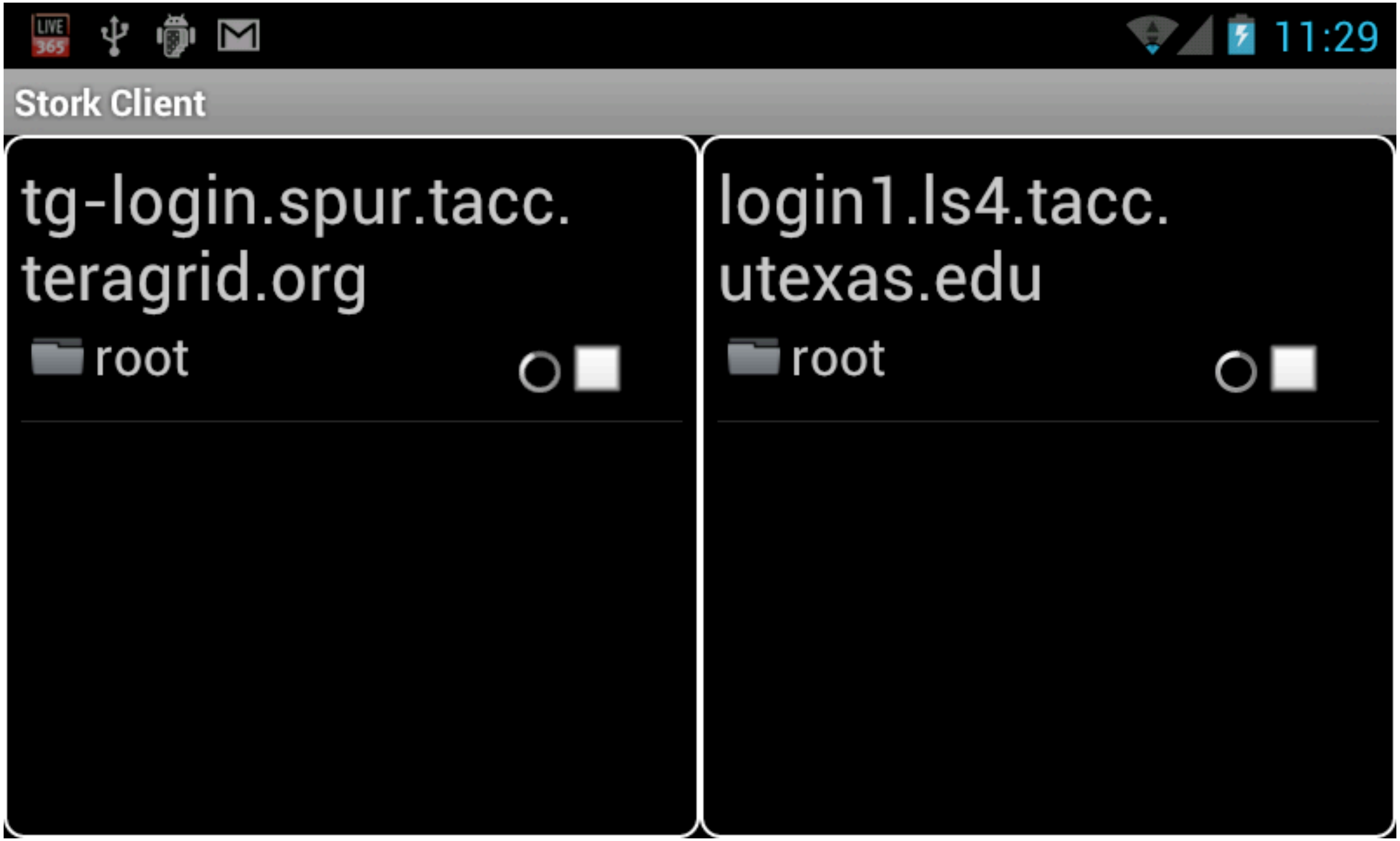
tg-login.spur.tacc.teragrid.org

root ○ □

login1.ls4.tacc.utexas.edu

root □

Stork Android Client



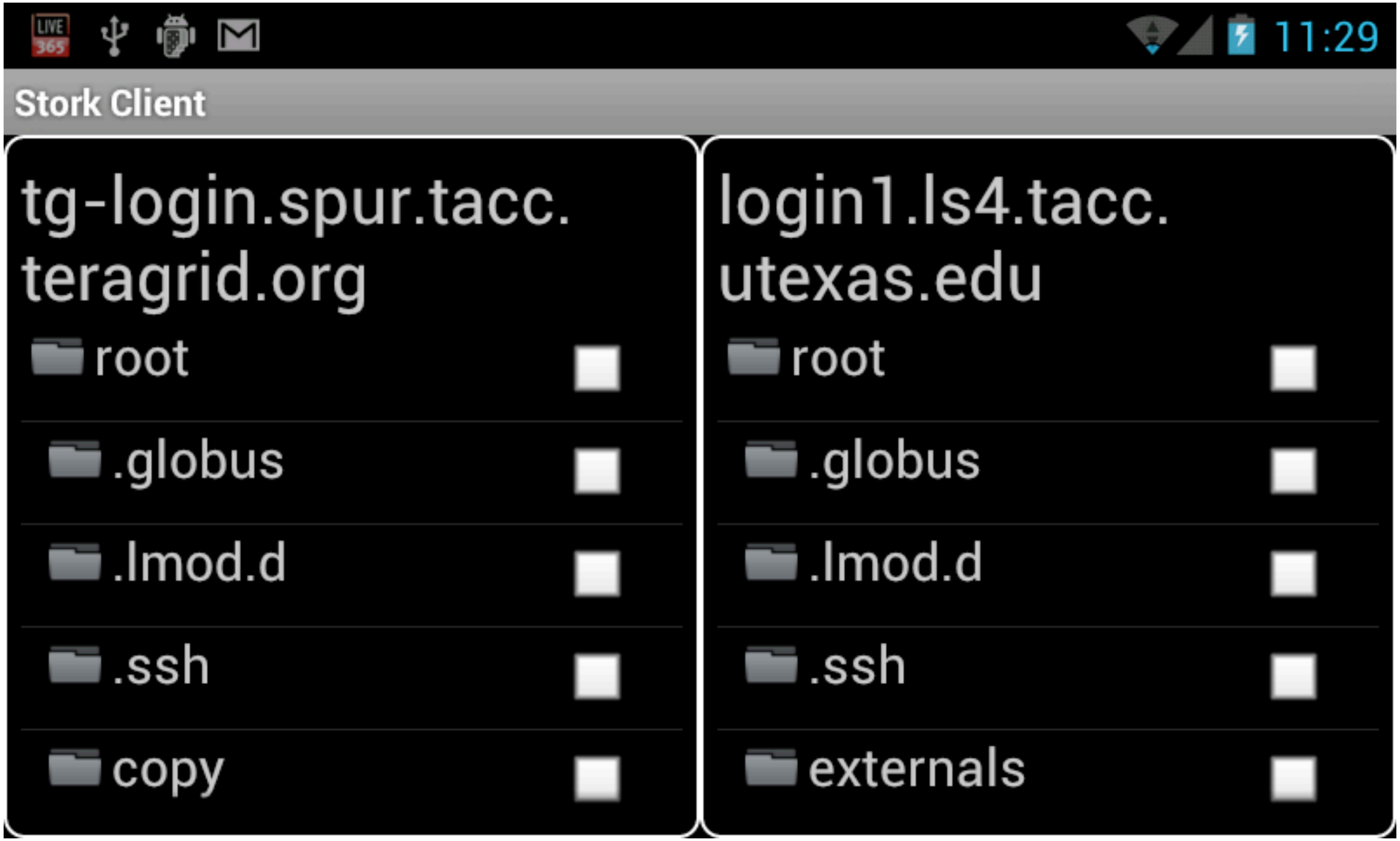
tg-login.spur.tacc.teragrid.org

root ○ □

login1.ls4.tacc.utexas.edu

root ○ □

Stork Android Client



Stork Client

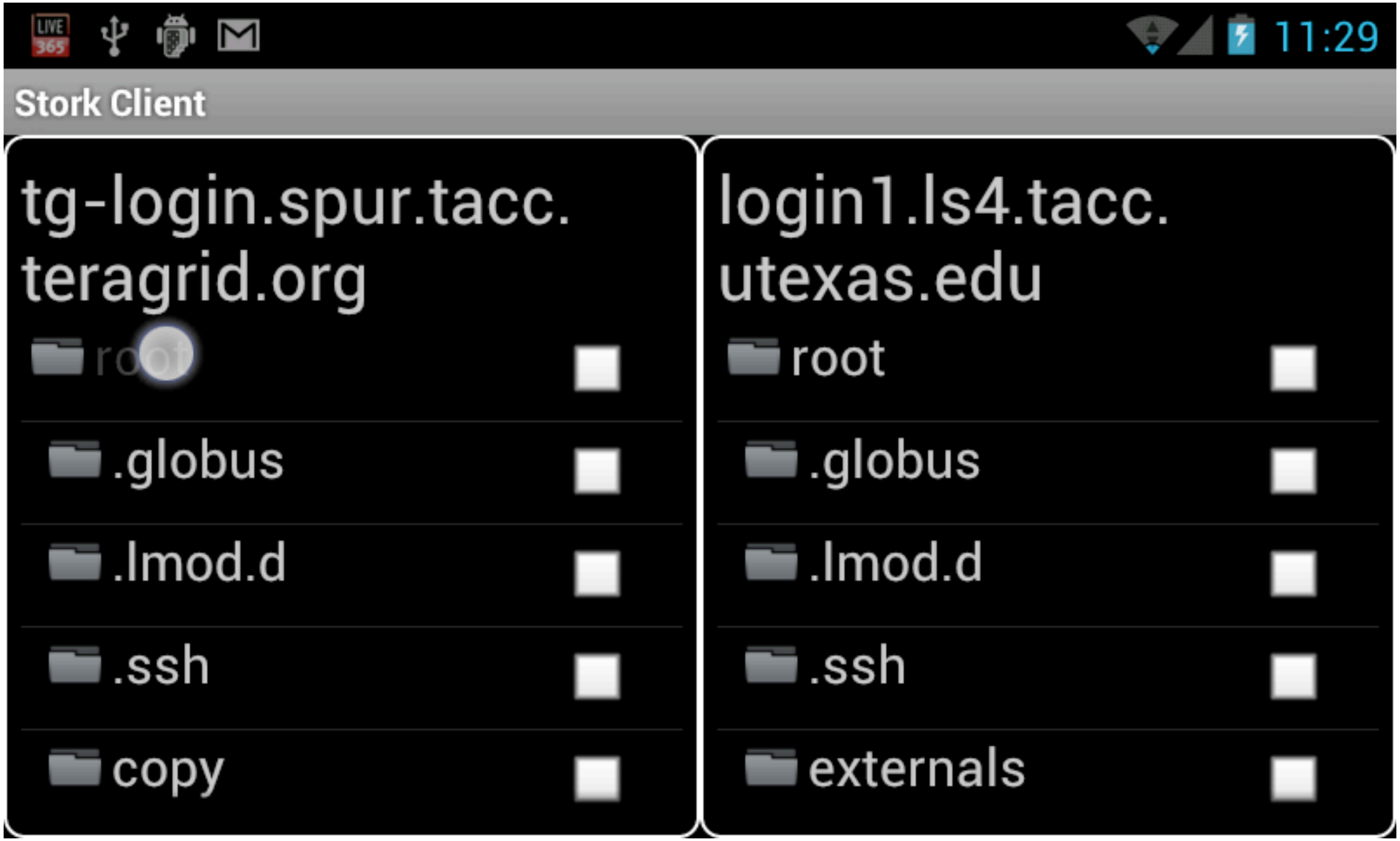
tg-login.spur.tacc.teragrid.org

- root □
- .globus □
- .lmod.d □
- .ssh □
- copy □

login1.ls4.tacc.utexas.edu

- root □
- .globus □
- .lmod.d □
- .ssh □
- externals □

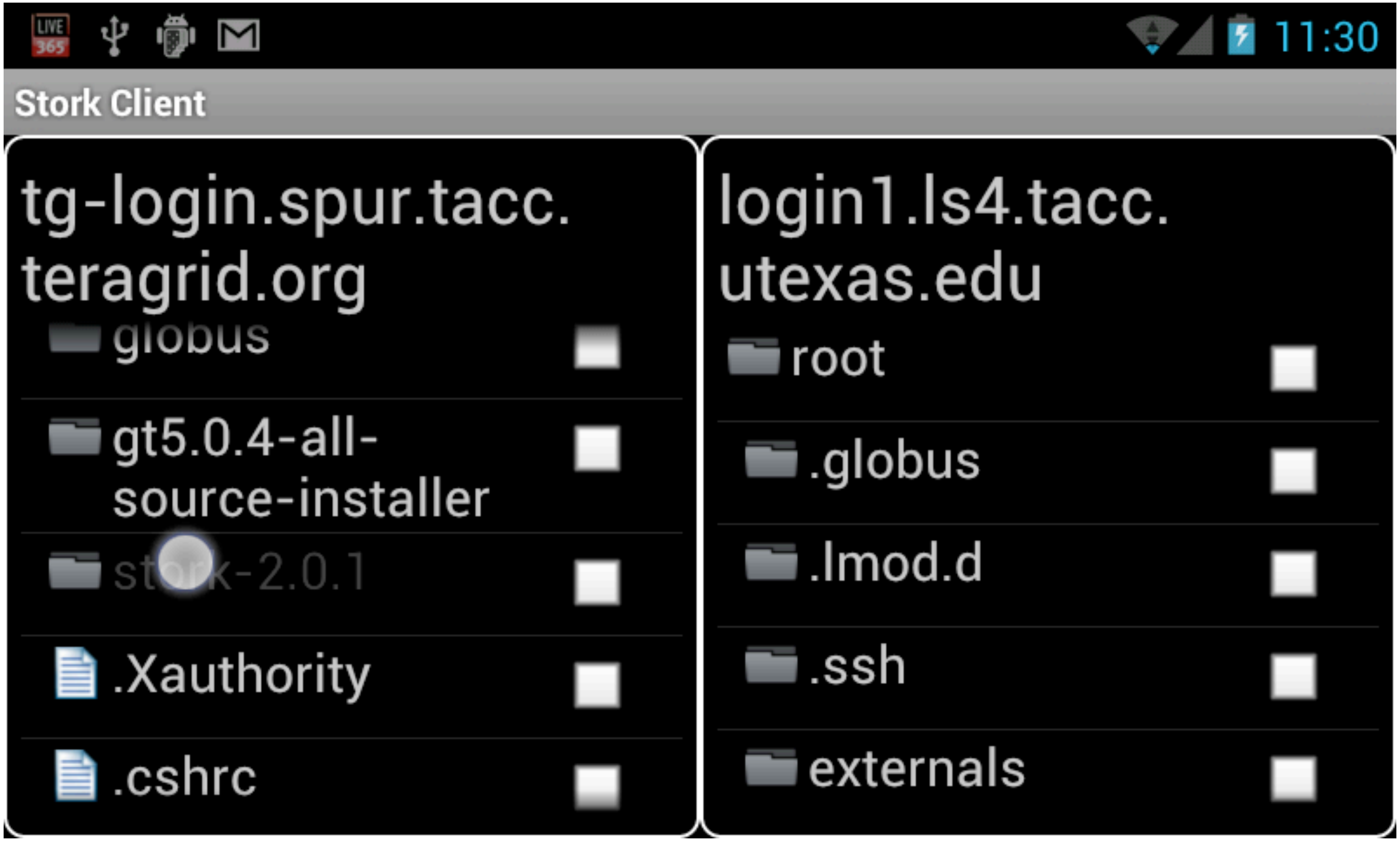
Stork Android Client



Stork Client

tg-login.spur.tacc.teragrid.org		login1.ls4.tacc.utexas.edu	
root	☐	root	☐
.globus	☐	.globus	☐
.lmod.d	☐	.lmod.d	☐
.ssh	☐	.ssh	☐
copy	☐	externals	☐

Stork Android Client



11:30

Stork Client

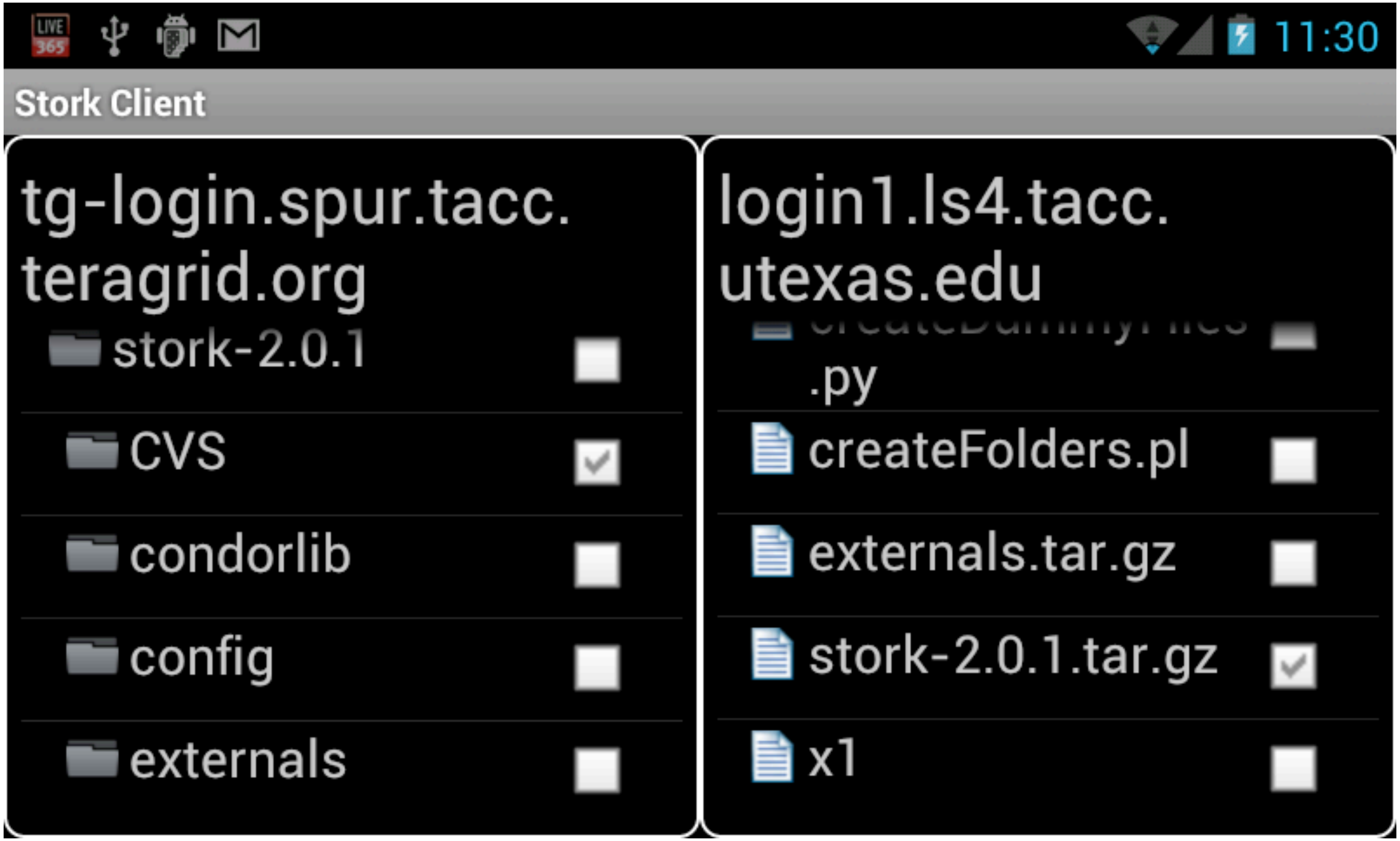
tg-login.spur.tacc.teragrid.org

- globus □
- gt5.0.4-all-source-installer □
- stork-2.0.1 □
- .Xauthority □
- .cshrc □

login1.ls4.tacc.utexas.edu

- root □
- .globus □
- .lmod.d □
- .ssh □
- externals □

Stork Android Client



Stork Client

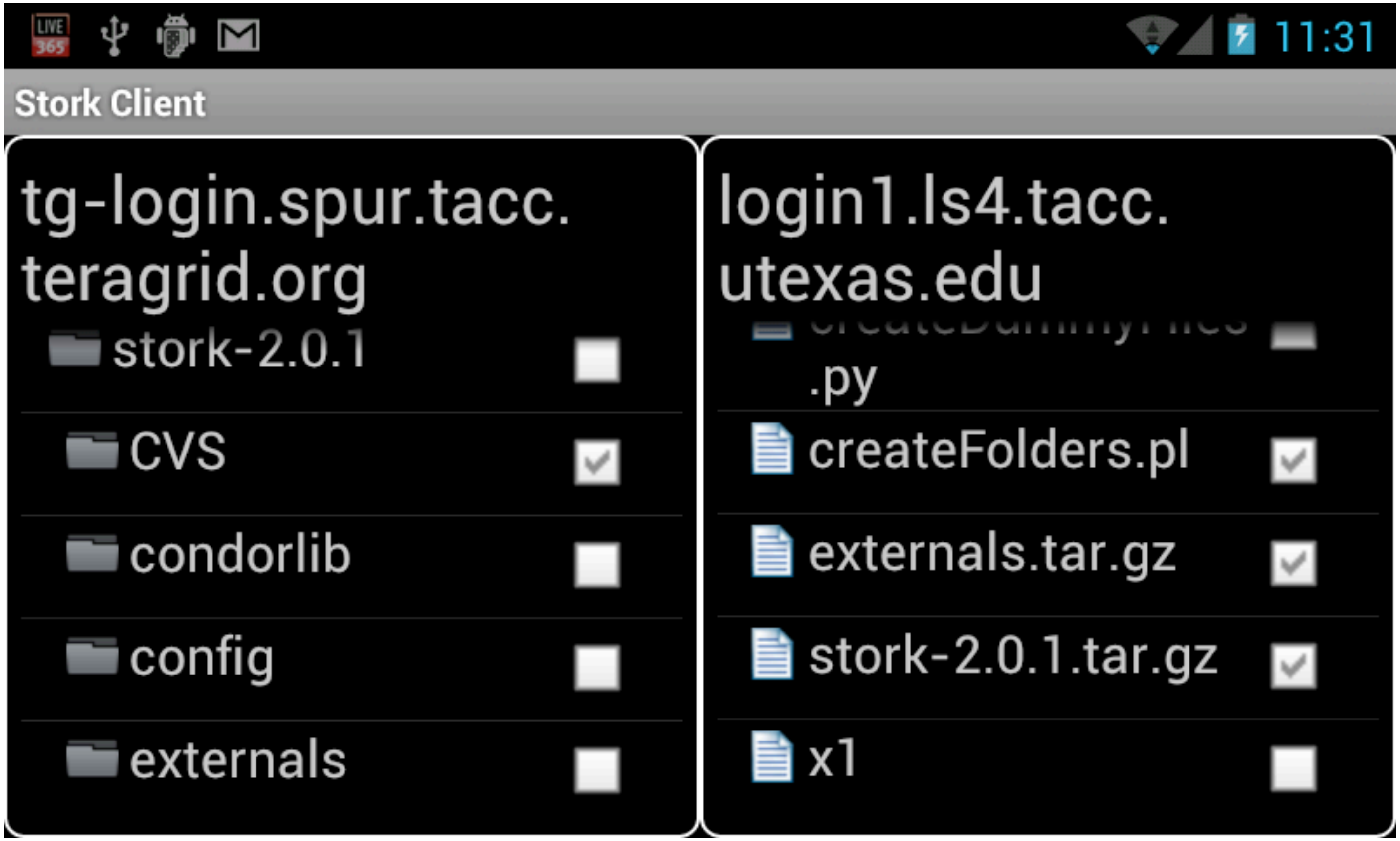
tg-login.spur.tacc.teragrid.org

- stork-2.0.1 □
- CVS ☑
- condorlib □
- config □
- externals □

login1.ls4.tacc.utexas.edu

- createDummyFiles.py □
- createFolders.pl □
- externals.tar.gz □
- stork-2.0.1.tar.gz ☑
- x1 □

Stork Android Client



Stork Client

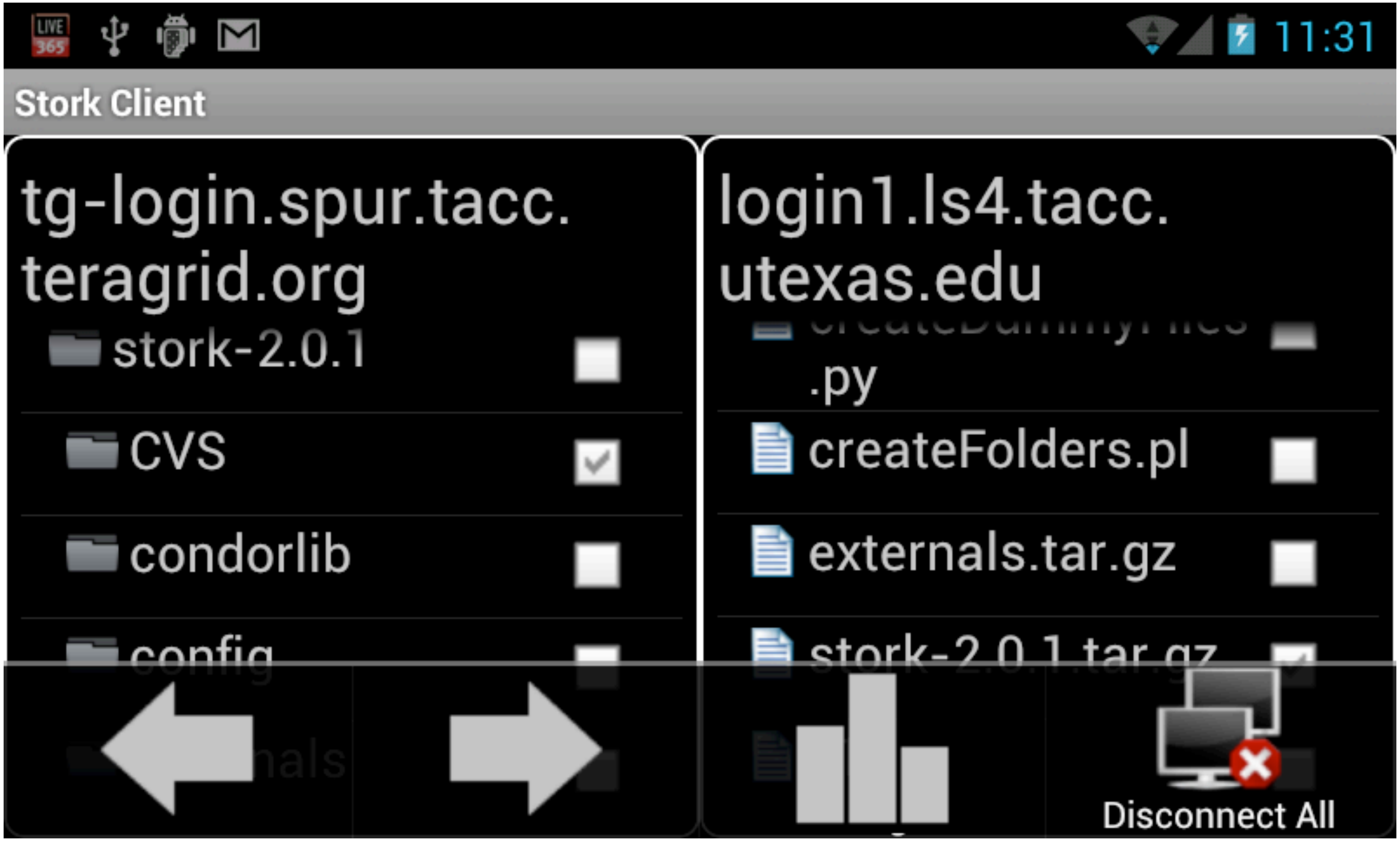
tg-login.spur.tacc.teragrid.org

- stork-2.0.1 □
- CVS ☑
- condorlib □
- config □
- externals □

login1.ls4.tacc.utexas.edu

- createDummyFiles.py
- createFolders.pl ☑
- externals.tar.gz ☑
- stork-2.0.1.tar.gz ☑
- x1 □

Stork Android Client



Stork Client

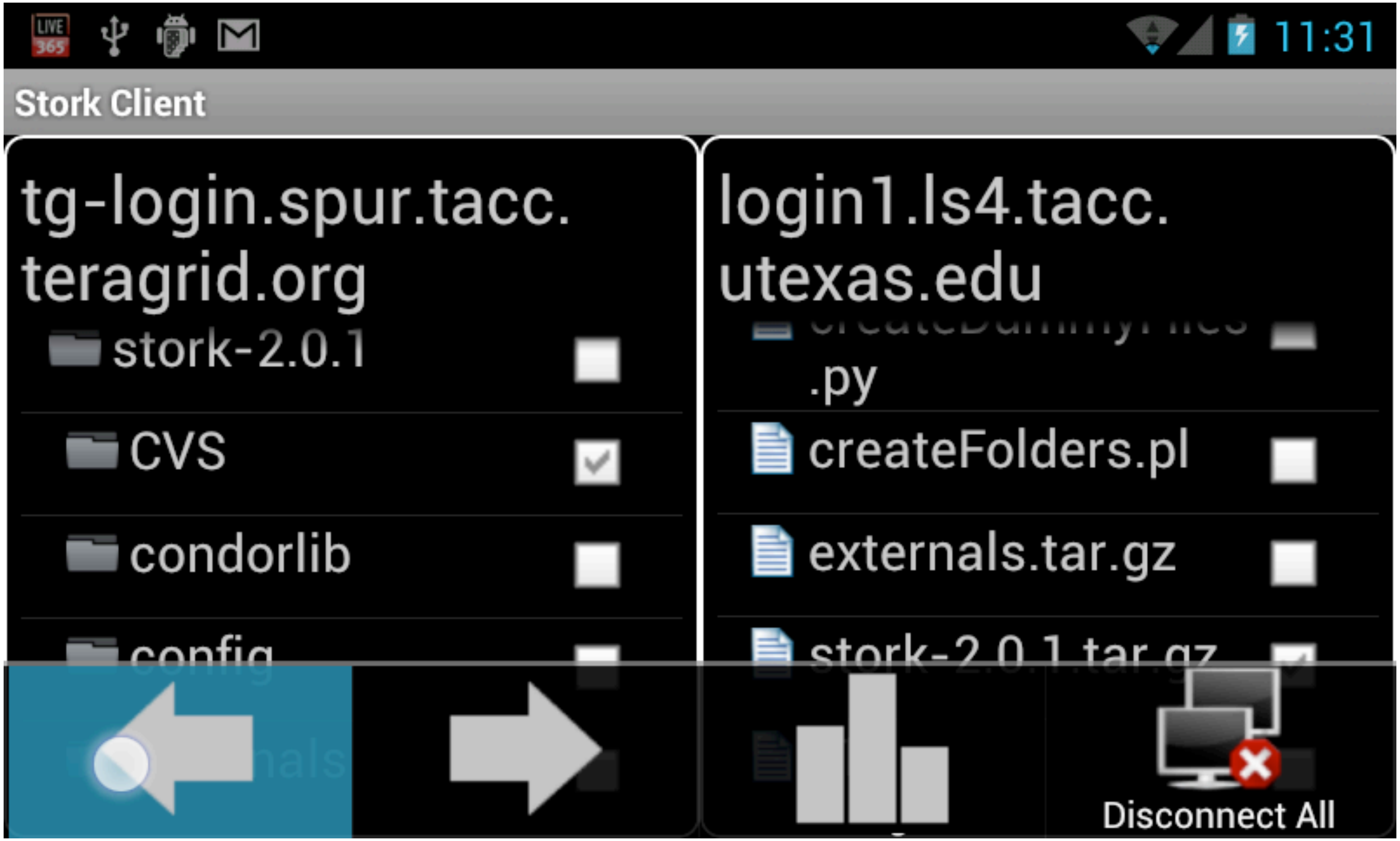
tg-login.spur.tacc.teragrid.org

- stork-2.0.1 □
- CVS ☑
- condorlib □
- config □

login1.ls4.tacc.utexas.edu

- createDummyFiles.py □
- createFolders.pl □
- externals.tar.gz □
- stork-2.0.1.tar.gz ☑

Stork Android Client



Stork Client

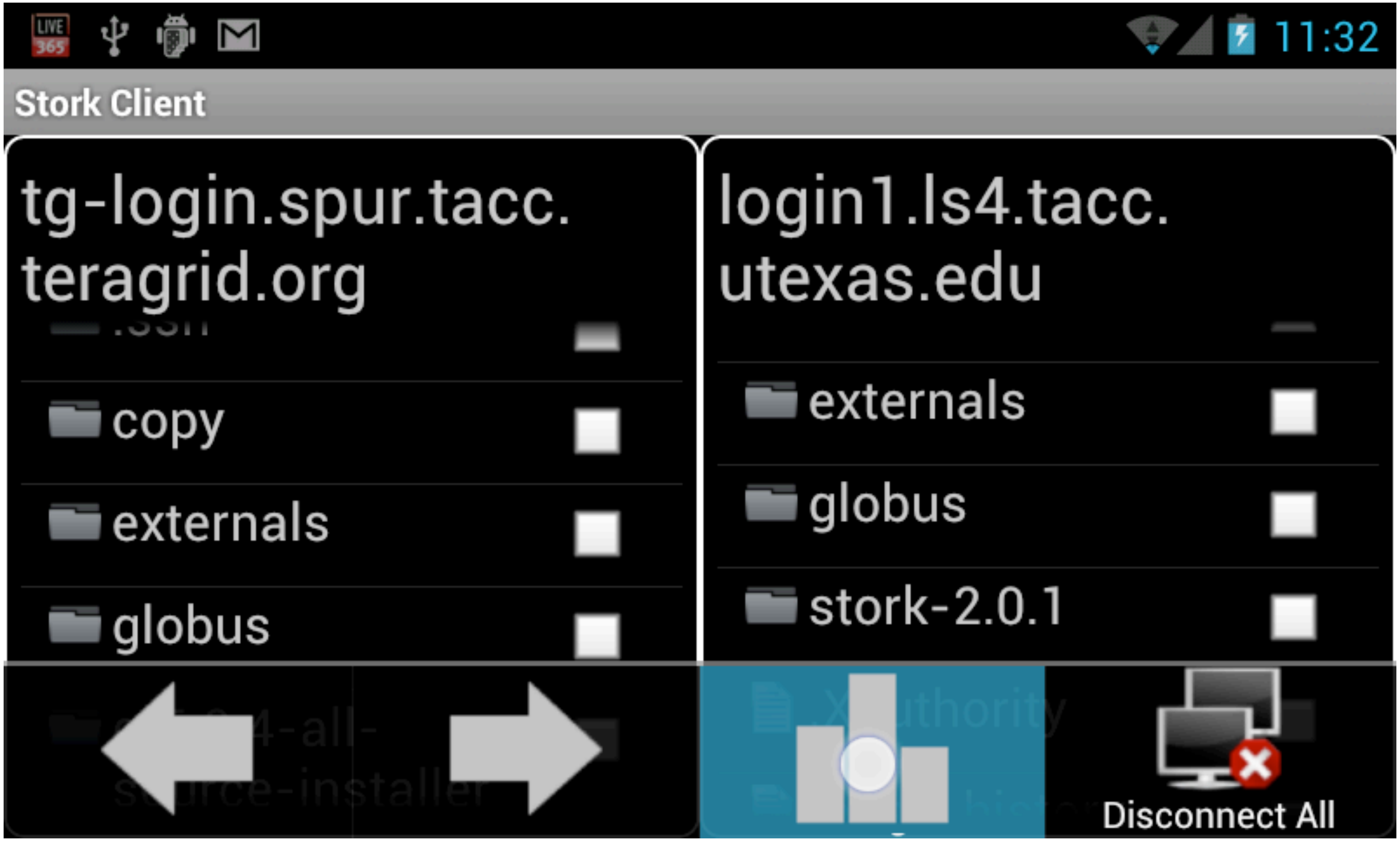
tg-login.spur.tacc.teragrid.org

- stork-2.0.1 □
- CVS ☑
- condorlib □
- config □

login1.ls4.tacc.utexas.edu

- createDummyFiles.py □
- createFolders.pl □
- externals.tar.gz □
- stork-2.0.1.tar.gz ☑

Stork Android Client



11:32

Stork Client

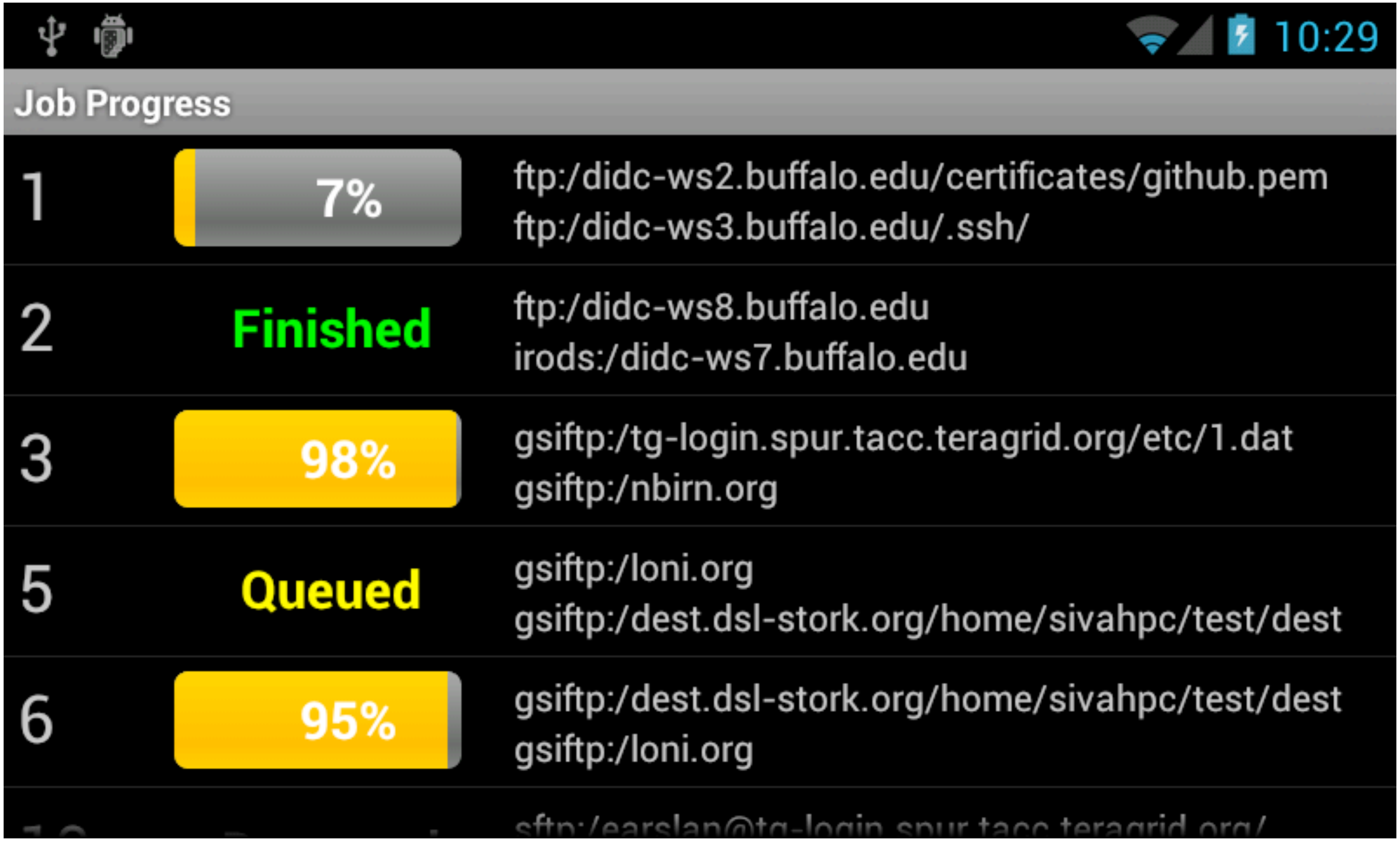
tg-login.spur.tacc.teragrid.org

- .ssh □
- copy □
- externals □
- globus □

login1.ls4.tacc.utexas.edu

- externals □
- globus □
- stork-2.0.1 □

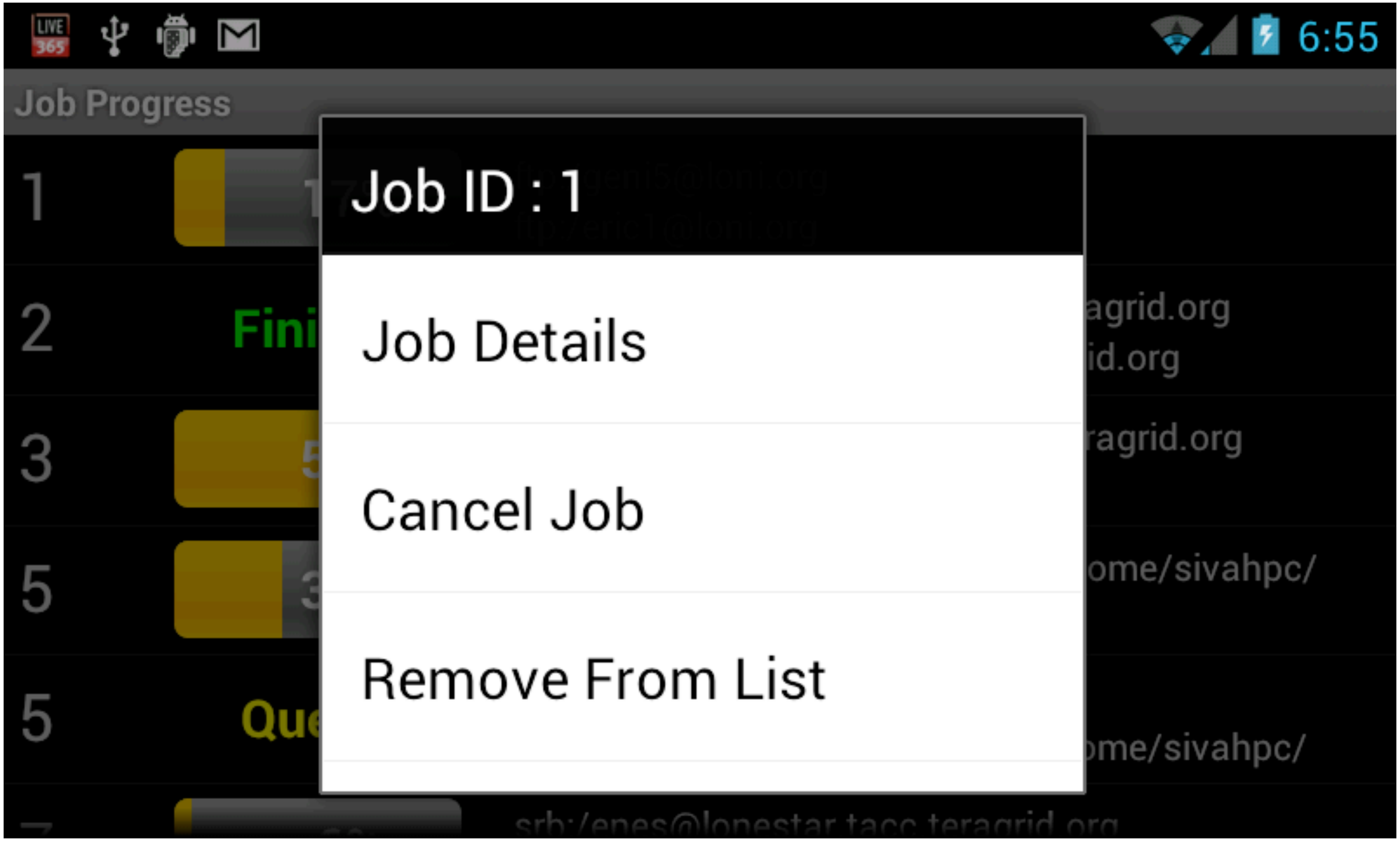
Stork Android Client



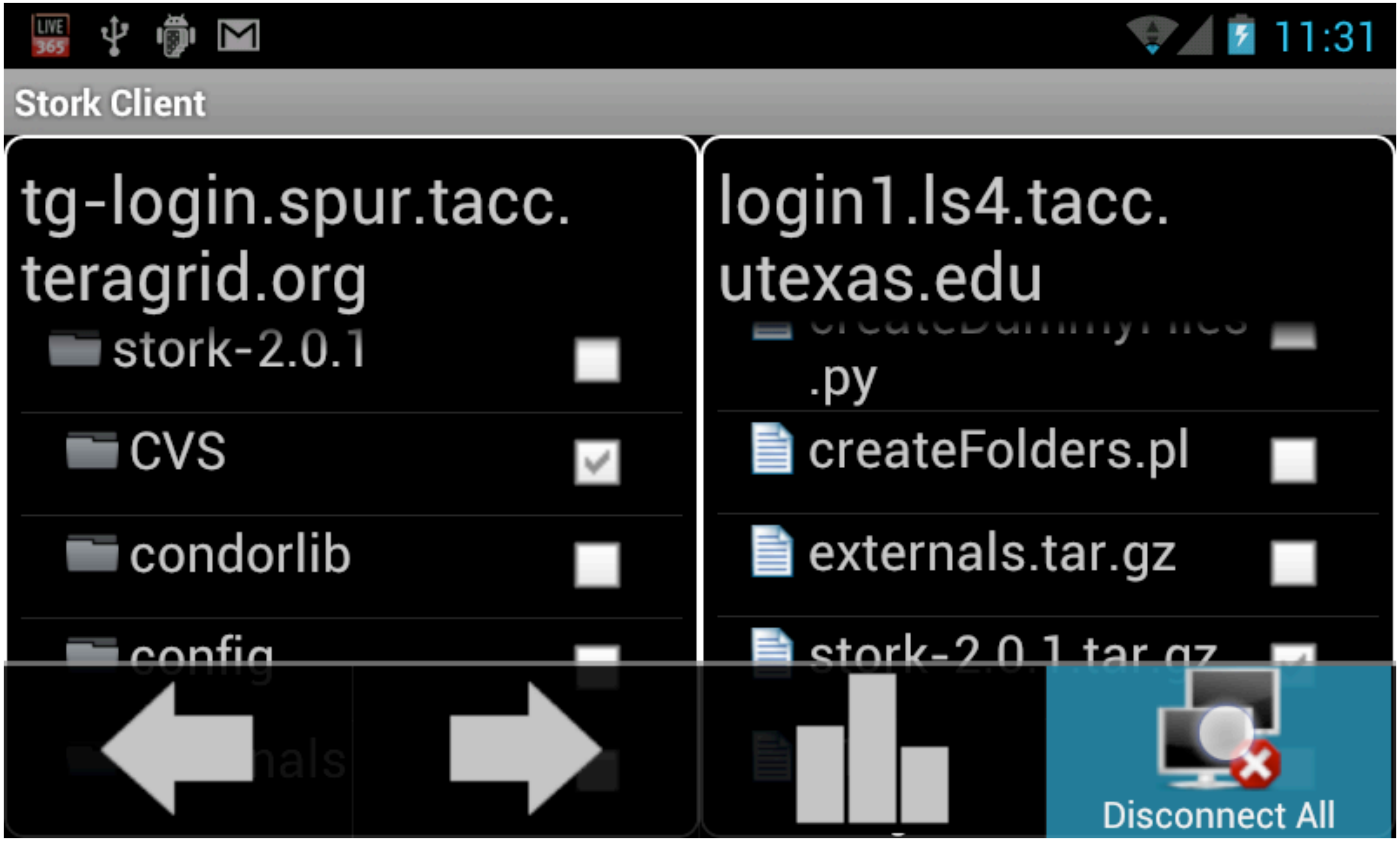
Job Progress

1	7%	ftp:/didc-ws2.buffalo.edu/certificates/github.pem ftp:/didc-ws3.buffalo.edu/.ssh/
2	Finished	ftp:/didc-ws8.buffalo.edu irods:/didc-ws7.buffalo.edu
3	98%	gsiftp:/tg-login.spur.tacc.teragrid.org/etc/1.dat gsiftp:/nbirn.org
5	Queued	gsiftp:/loni.org gsiftp:/dest.dsl-stork.org/home/sivahpc/test/dest
6	95%	gsiftp:/dest.dsl-stork.org/home/sivahpc/test/dest gsiftp:/loni.org
		sftp:/earslan@tg-login.spur.tacc.teragrid.org/

Stork Android Client



Stork Android Client



11:31

Stork Client

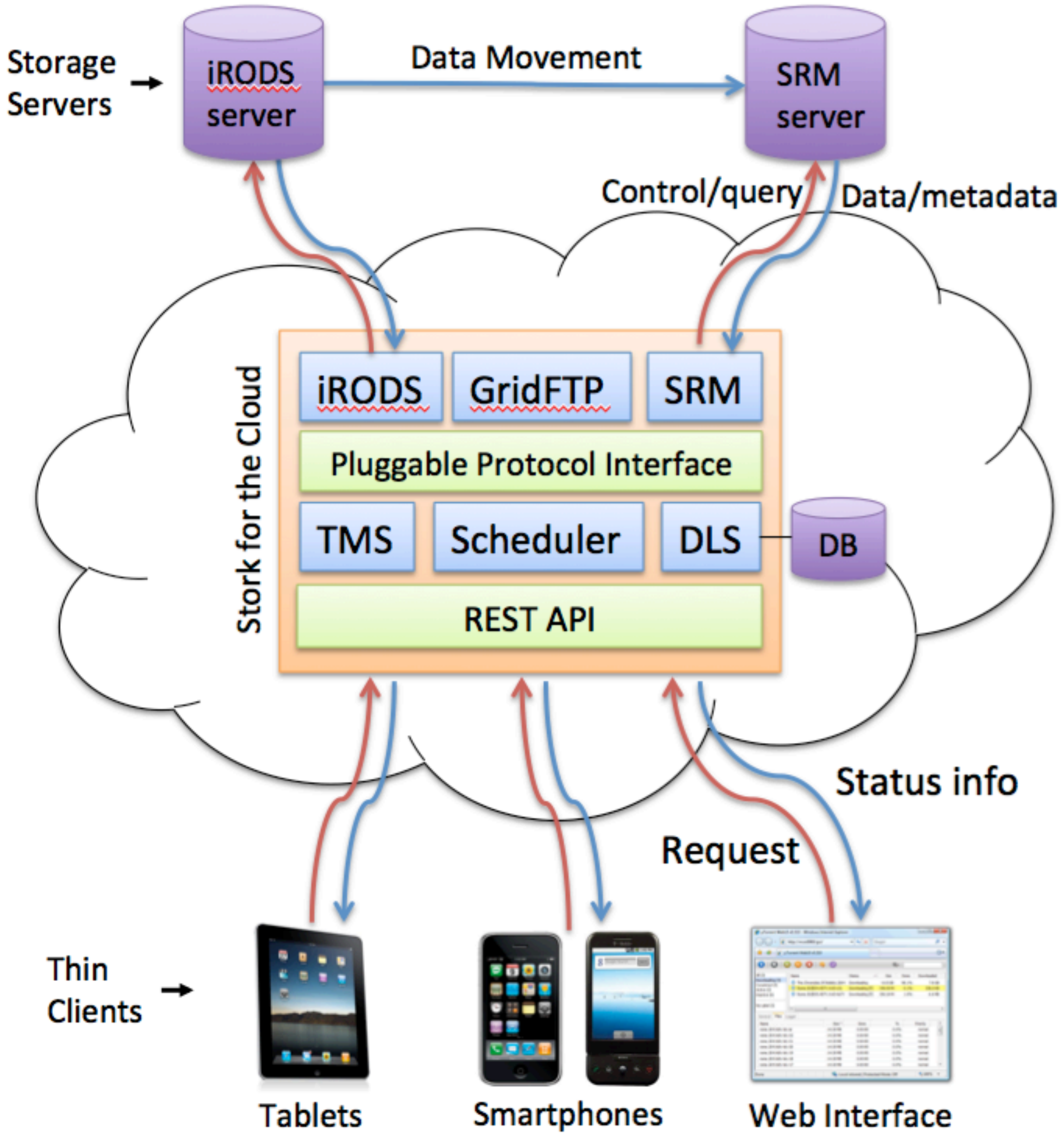
tg-login.spur.tacc.teragrid.org

- stork-2.0.1 □
- CVS ☑
- condorlib □
- config □

login1.ls4.tacc.utexas.edu

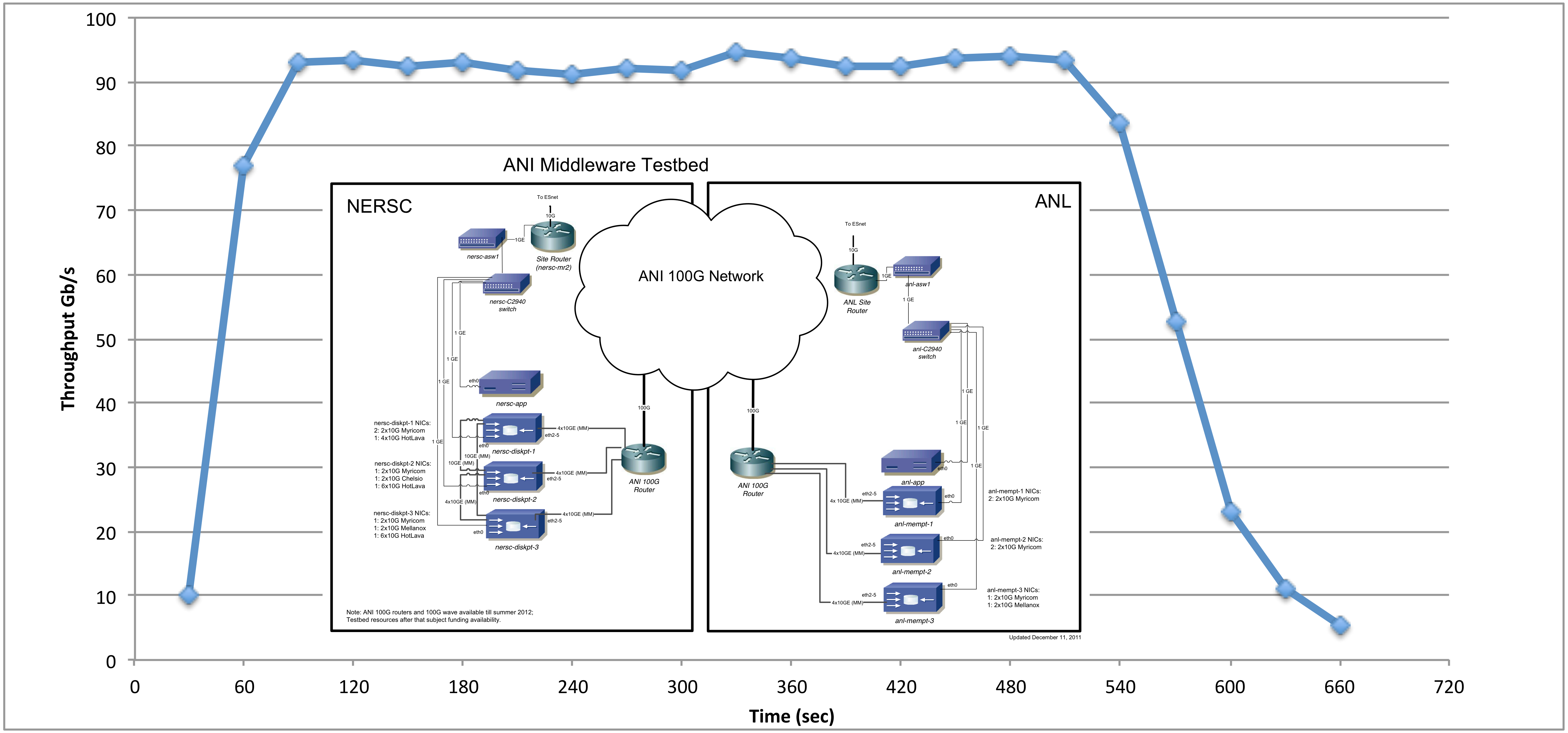
- createDummyFiles.py □
- createFolders.pl □
- externals.tar.gz □
- stork-2.0.1.tar.gz ☑

Stork for the Cloud



- Prototype implementation complete, testing stage
- Will be deployed as hosted service
- Allow deployment on private clouds as well
- Available on Amazon EC2 and Windows Azure
- More optimizations coming

100 Gbit Performance



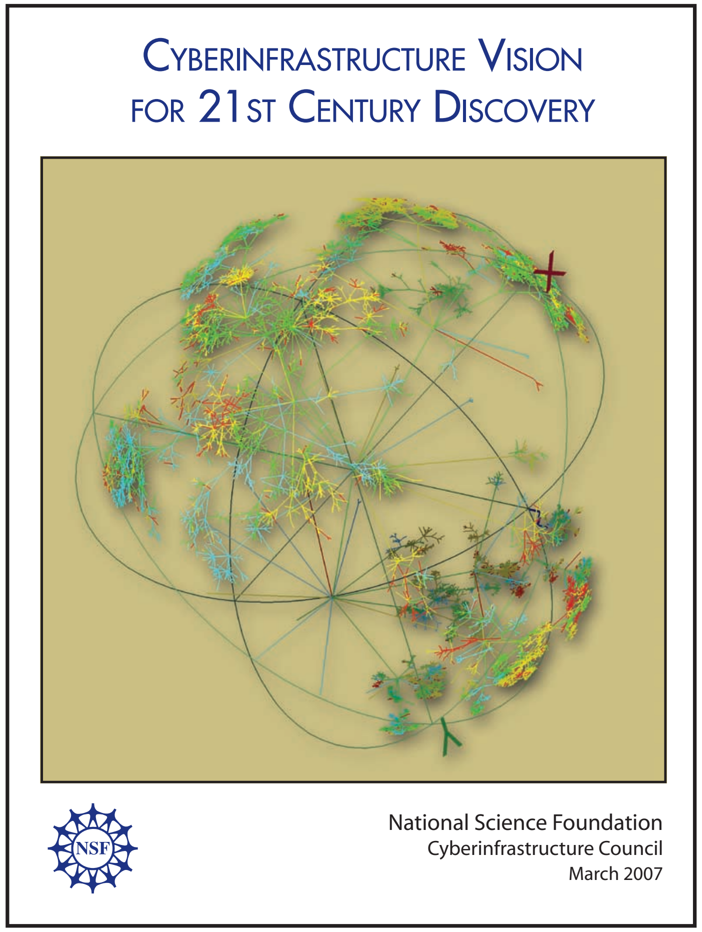
ANI Middleware Testbed

Note: ANI 100G routers and 100G wave available till summer 2012;
Testbed resources after that subject funding availability.

Updated December 11, 2011

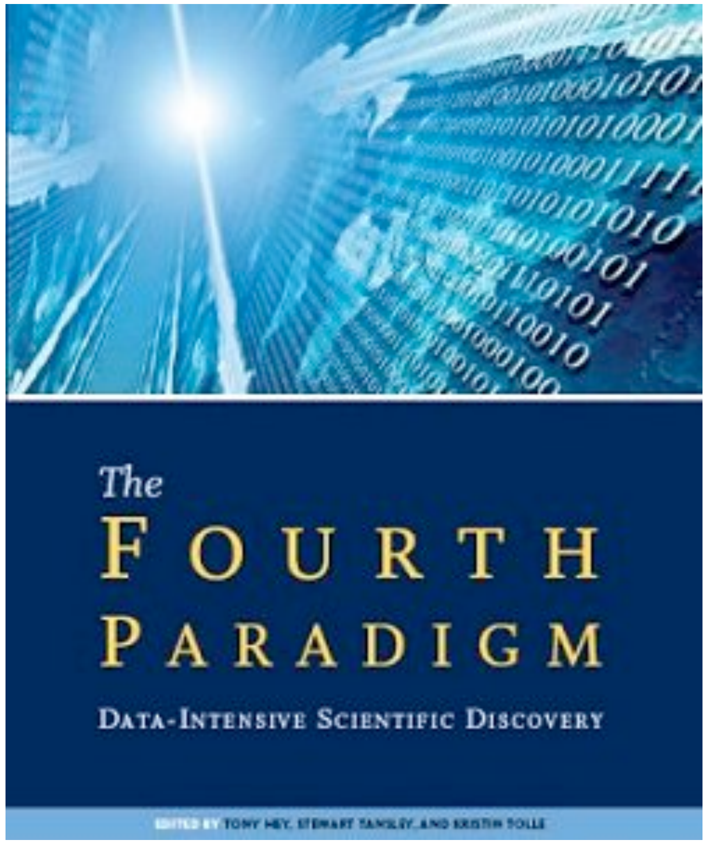
Summary

- Scientific and commercial applications are getting more and more data intensive
- Data sharing and bulk data transfers are still a major bottleneck in front of multi-institutional and inter-disciplinary collaborative science
- Stork for the Cloud provides end-to-end throughput optimization in hosted environment accessible through ultra-thin clients



CYBERINFRASTRUCTURE VISION
FOR 21ST CENTURY DISCOVERY

National Science Foundation
Cyberinfrastructure Council
March 2007



The
FOURTH
PARADIGM

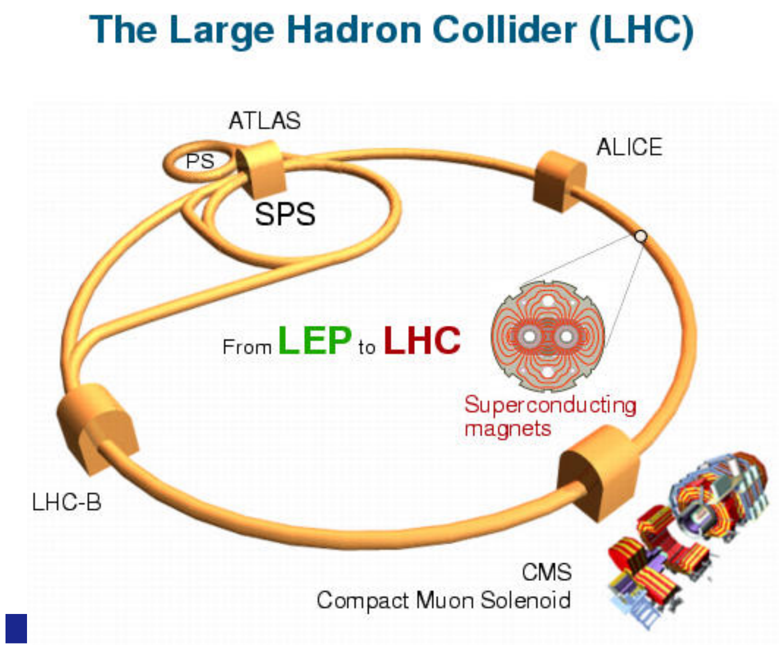
DATA-INTENSIVE SCIENTIFIC DISCOVERY

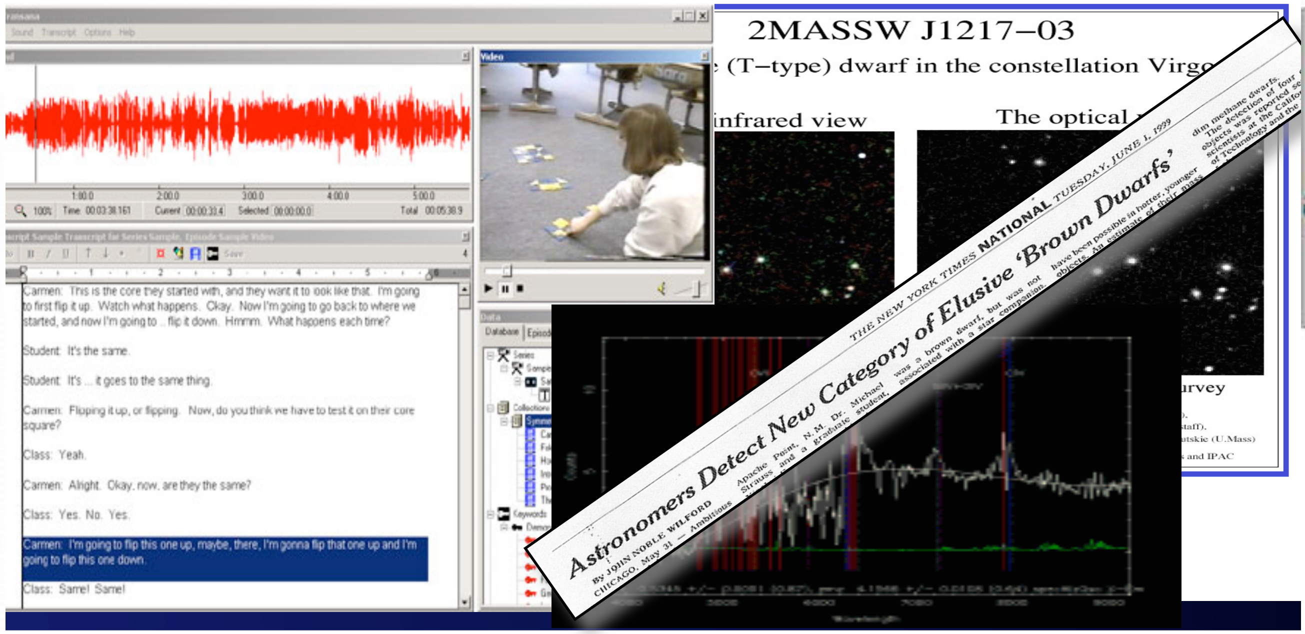
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

This work has been sponsored by:
NSF, DOE, ONR, NOAA

For more information:

Stork web page: http://www.storkproject.org





Time 00:03:38.161 Current 00:00:33.4 Selected 00:00:00.0 Total 00:05:38.9

Carmen: This is the core they started with, and they want it to look like that. I'm going to first flip it up. Watch what happens. Okay. Now I'm going to go back to where we started, and now I'm going to .. flip it down. Hmmm. What happens each time?

Student: It's the same.

Student: It's ... it goes to the same thing.

Carmen: Floping it up, or flipping. Now, do you think we have to test it on their core square?

Class: Yeah.

Carmen: Alright. Okay, now, are they the same?

Class: Yes. No. Yes.

Carmen: I'm going to flip this one up, maybe, there, I'm gonna flip that one up and I'm going to flip this one down.

Class: Same! Same!

2MASSW J1217–03

(T–type) dwarf in the constellation Virgo

infrared view

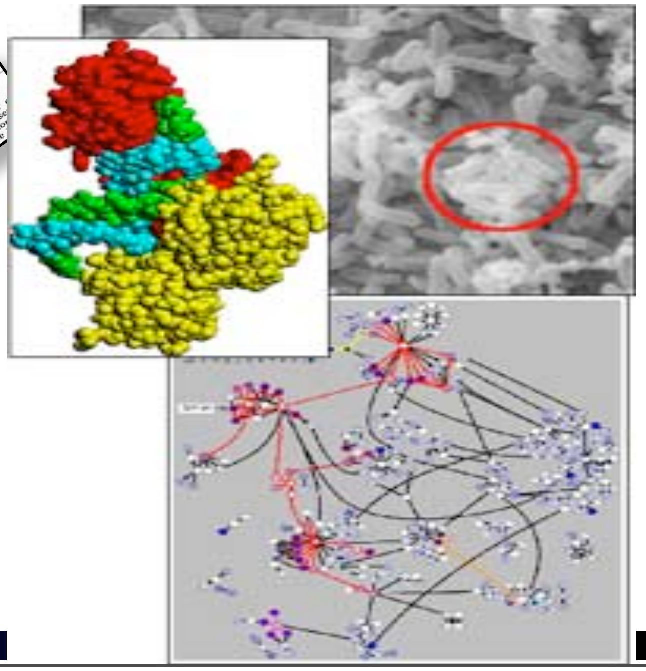
The optical

THE NEW YORK TIMES NATIONAL TUESDAY, JUNE 1, 1999

Astronomers Detect New Category of Elusive 'Brown Dwarfs'

By JOHN NOBLE WILFORD

CHICAGO, May 31 — Ambitious

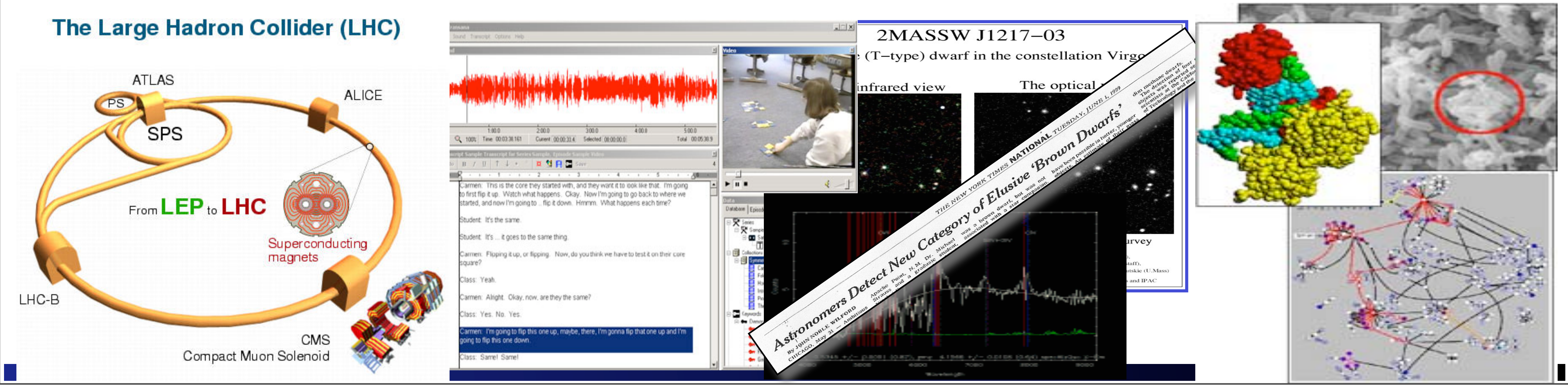


This work has been sponsored by:

NSF, DOE, ONR, NOAA

For more information:

Stork web page: http://www.storkproject.org



Carmen: This is the core they started with, and they want it to look like that. I'm going to first flip it up. Watch what happens. Okay. Now I'm going to go back to where we started, and now I'm going to .. flip it down. Hmmm. What happens each time?

Student: It's the same.

Student: It's ... it goes to the same thing.

Carmen: Floping it up, or flipping. Now, do you think we have to test it on their core square?

Class: Yeah.

Carmen: Alright. Okay, now, are they the same?

Class: Yes. No. Yes.

Carmen: I'm going to flip this one up, maybe, there, I'm gonna flip that one up and I'm going to flip this one down.

Class: Same! Same!

THE NEW YORK TIMES NATIONAL TUESDAY, JUNE 1, 1999

Astronomers Detect New Category of Elusive 'Brown Dwarfs'

By JOHN NOBLE WILFORD

CHICAGO, May 31 — Ambitious Apache Point, N.M. Dr. Michael Strauss and a graduate student, was a brown dwarf, but was not associated with a star companion. have been possible in hotter, younger objects. An estimate of their mass dim methane dwarfs. The detection of four objects was reported scientists at the Califo of Technology and the