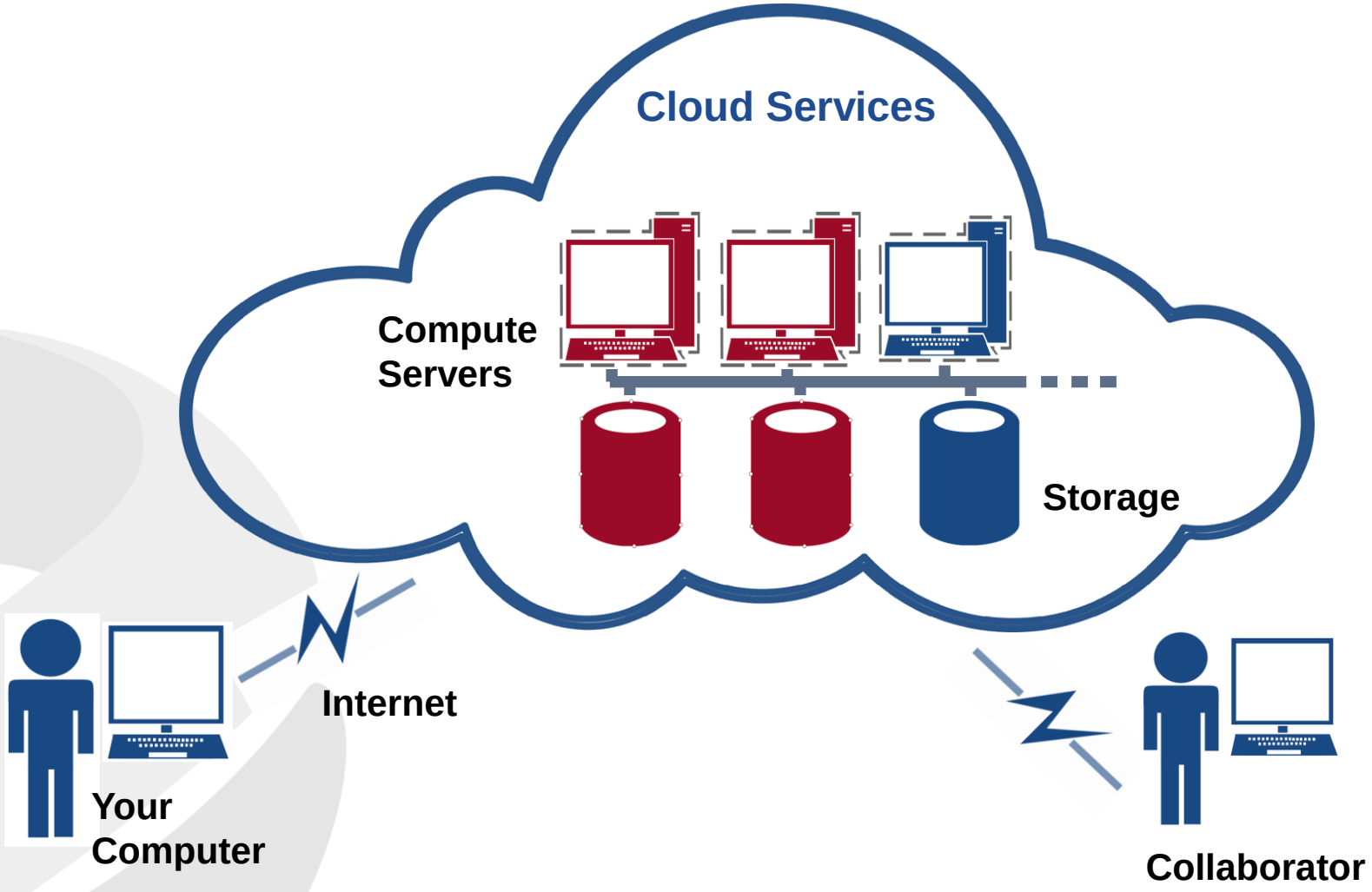
WORKSHOP:
Running a Virtual Machine in the Cloud

Joey Gerlach

25 February 2016

WORKSHOP: Running a Virtual Machine in the Cloud

WHAT IS CLOUD COMPUTING?


Cloud Services
Compute Servers
Storage
Internet
Your Computer
Collaborator

Nectar Services

- 2010: Australian research community voices their need for flexible, low cost, on demand computing resources.
- The Australian Research Cloud has been funded by the Commonwealth Government of Australia through the NeCTAR project (National eResearch Collaboration Tools and Resources)
- It is sustained by participating universities, research institutions and State governments

What is cloud computing

Simply put, cloud computing means:

Storing and accessing **data and programs** *over the Internet* instead of on your own computers hard drive.

Your local computer is only used to connect to and control the resources.

Cloud Computing

Cloud computing enables IT infrastructure to be:

- More flexible:
  - Re-provision resources according to your needs.
- Easier to use:
  - Provision resources quickly and easily; Little maintenance.
- Cheaper:
  - Save significant expenditures for local infrastructure.

Benefits for your research

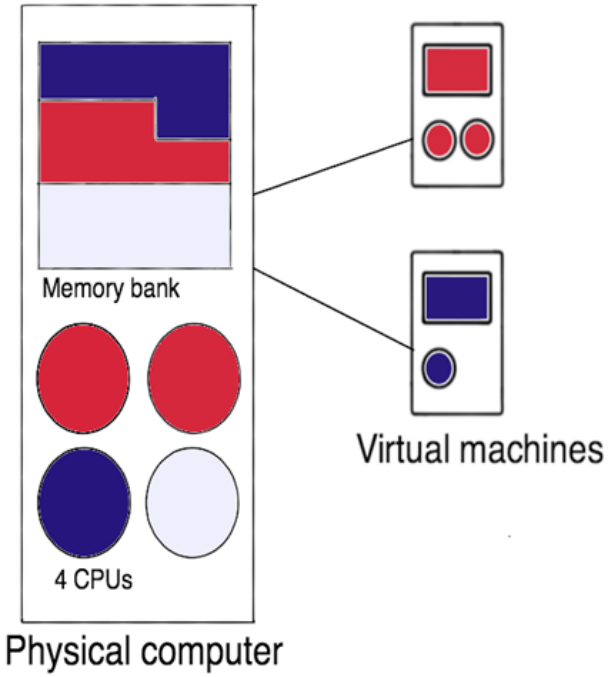
- Concentrate on your work instead of spending time to obtain and maintain hardware.
- Take advantage of a **cost-effective** IT infrastructure.
- Shared infrastructure → easier collaboration.

What the future holds

- Cloud computing is becoming a standard technology.
- Most efficient and easiest way to gain access to IT resources.
- “Sustainable research” through shared infrastructure

Virtualization

- **Virtualization** basically means that the hardware is “simulated”.
- We can simulate a whole computer including the Operating System.
- Several virtual computers can run on one larger, more powerful computer



Cloud *vs* HPC

- High Performance Computing (HPC) is not the same as cloud computing.
- Both technologies differ in a number of ways, and have some similarities as well.
- We may refer to both types as “large scale computing”.
  - Both systems target computing scalability differently.

High Performance Computing (HPC)
e.g. Tizard

- HPC targets extremely large sets of data and crunching the information in parallel, while sharing the data between compute nodes.
  - The data connection between the nodes has to be very fast.
  - The entire grid of nodes is turned into a single “supercomputer”.

- ***One*** application can be run ***across*** a variable number of nodes. We call this **vertical scalability.**

Cloud Computing

- Cloud computing targets “embarrassingly parallel problems” (EPP).
- The individual computers don’t have to be super fast.
- The power lies in having a huge number of computers.
- ***Several*** applications run on ***several*** nodes. We call this **horizontal scalability**.

When to use the Cloud

- You want **instant availability** of large-scale computing resources.
- Possibility **of software choice**: design virtual machines to suit your need, incl. choice of OS.
- The simple case: you need **easy access to computing infrastructure.**

The Cloud: Drawbacks

- **Requires Internet** to access—if it drops out, you lose access.
- **Indirect access control**: The ISPs and telecommunication companies control your Internet access.
- **Service outage** at the cloud service provider can take out your resources.
- Concerns about **ownership**: Who owns the data you store online?
- **Service charge** is based upon usage.

The Cloud: Advantages

- **Cost savings**: Nectar resources are free; building and maintaining on-premises infrastructure is expensive.
- **Individual setup**: Users can set up their own server.
- **Access independence**: Via the Internet from anywhere.
- **Large computing capacity** access quickly and easily.

The Cloud: Advantages

- **“Elasticity”** (Flexibility and Scalability): users can scale up or down resources as required at the time.
- **Resource sharing**: Multiple users can work on the same data.
- **Security** of professionally run data centres is often as good as, or better than maintaining local infrastructure.

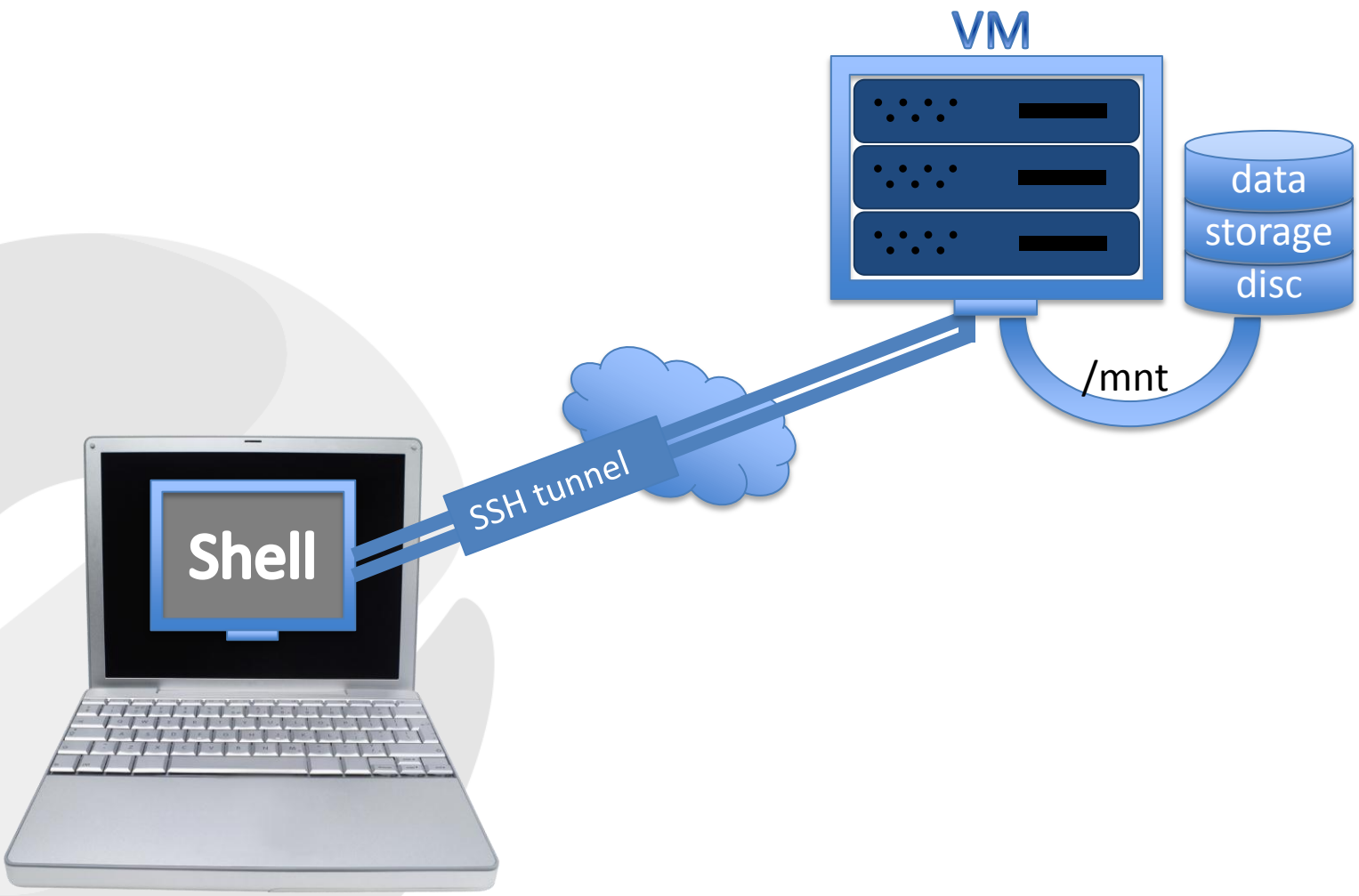
Services provided by the Cloud

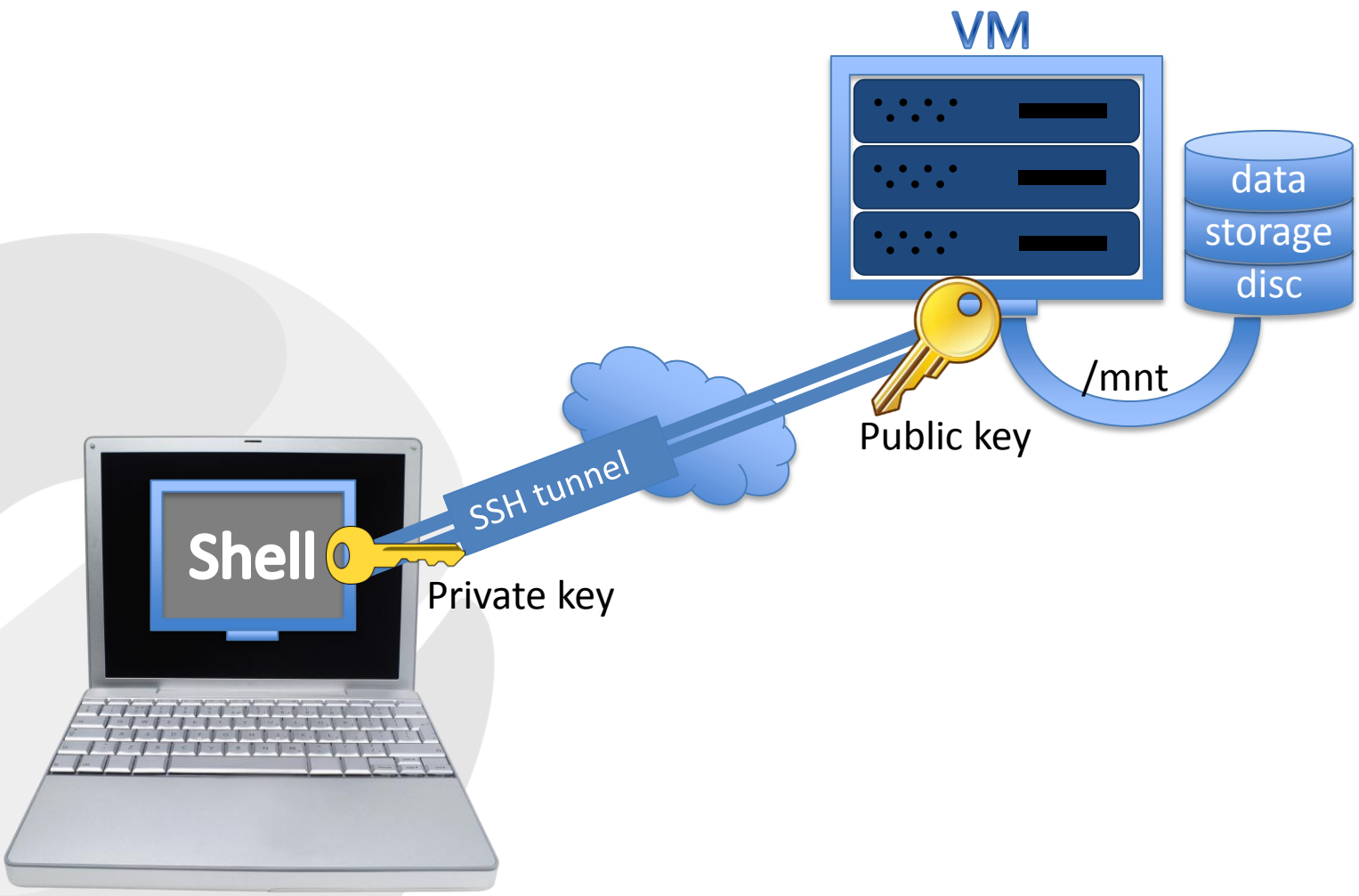
- Compute resources for running simulation and data analysis software
- Hosting of web sites, databases, web applications and other online resources
- Customised access to online data sets and data analysis tools.
- Virtualisation and hosting of virtualised servers.
- Hosting of online research tools and domain-specific virtual laboratories.

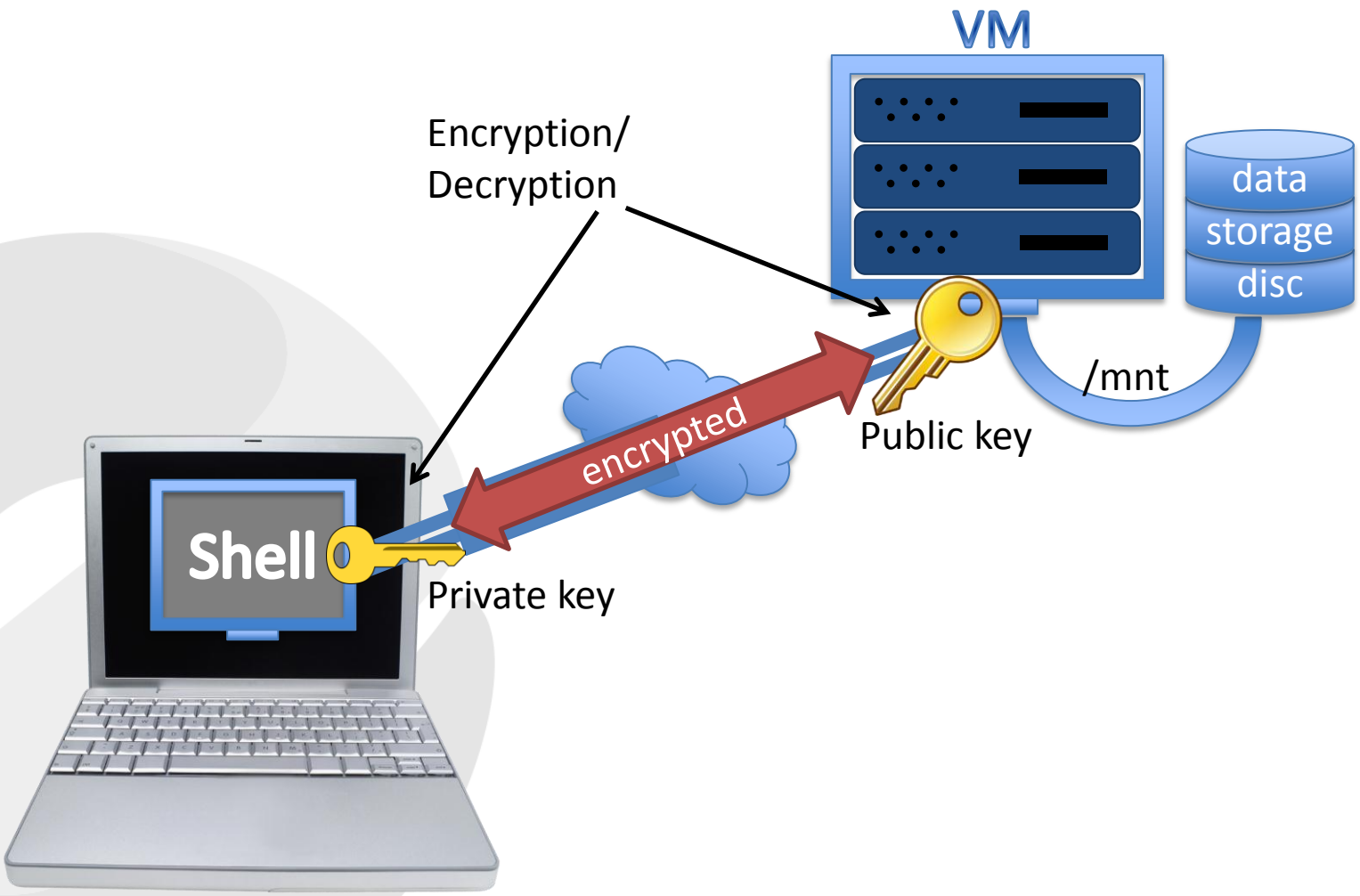
WORKSHOP: Running a Virtual Machine in the Cloud

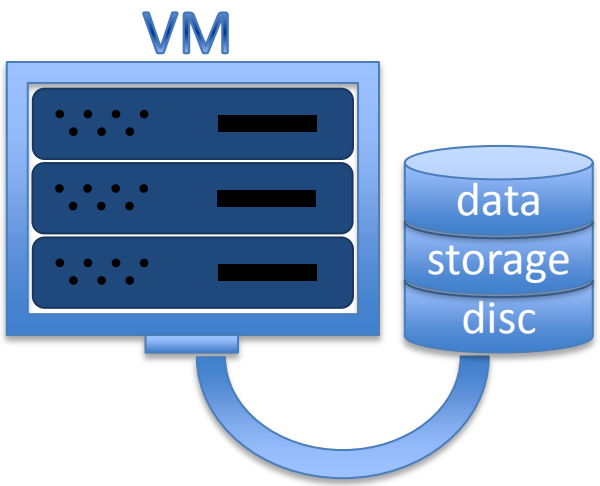
THE STRUCTURE OF THE VM

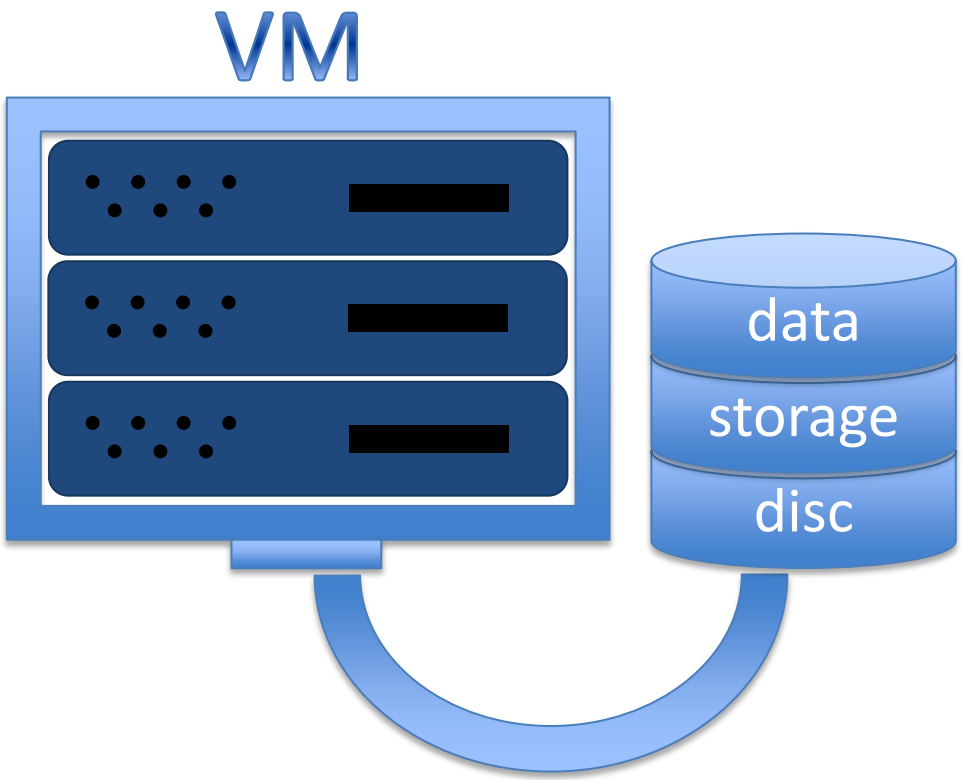
- Discs
- File Structure
- Users


VM
data
storage
disc
/mnt
SSH tunnel
Shell

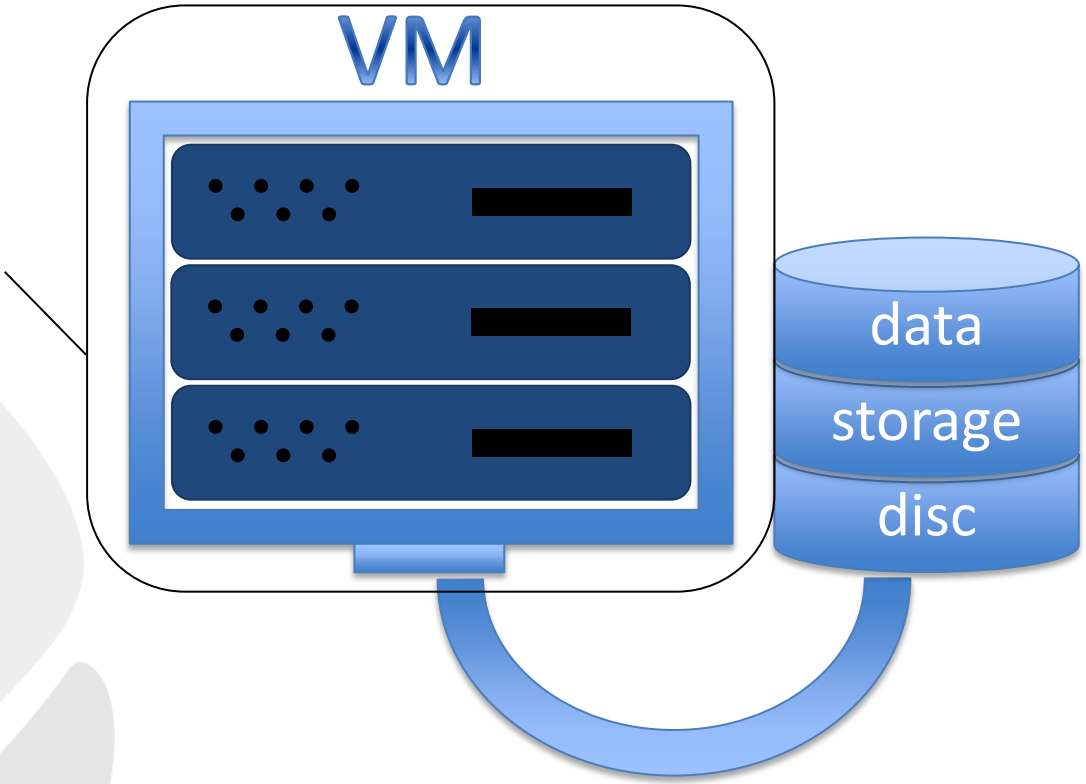

VM
data
storage
disc
/mnt
Public key
SSH tunnel
Shell
Private key


VM
Encryption/
Decryption
data
storage
disc
/mnt
Public key
encrypted
Shell
Private key


VM
data
storage
disc

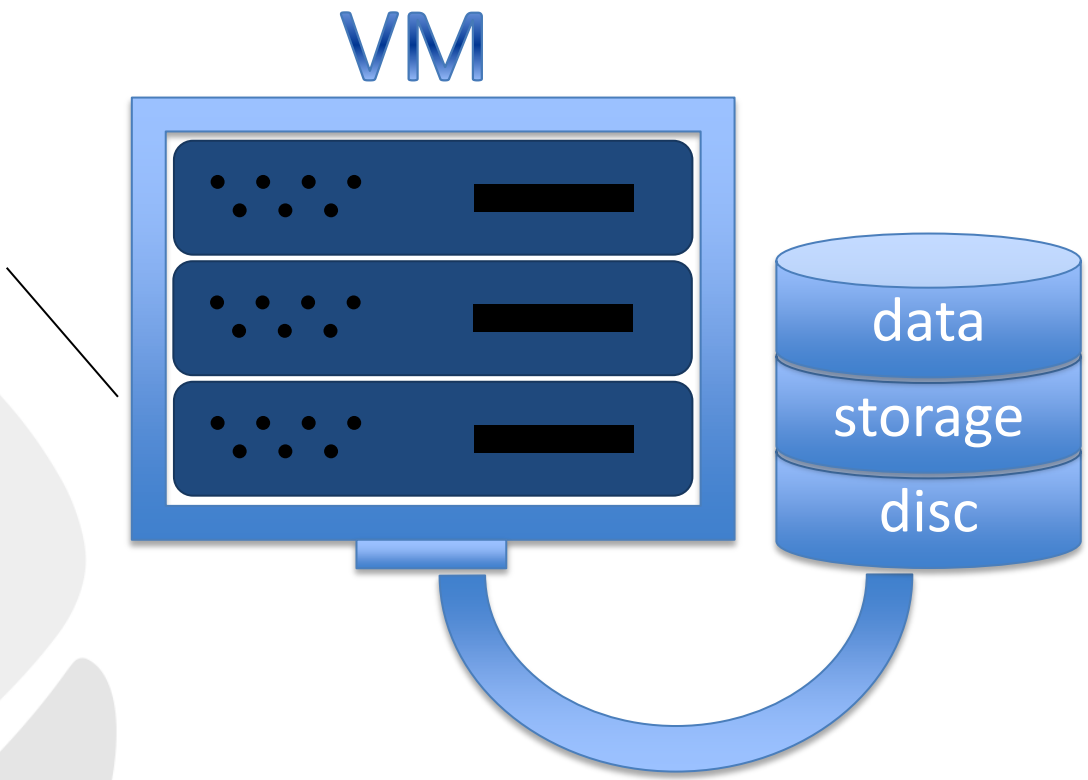

VM
data
storage
disc

/dev/vda
Root (primary) disc
5-30 GB



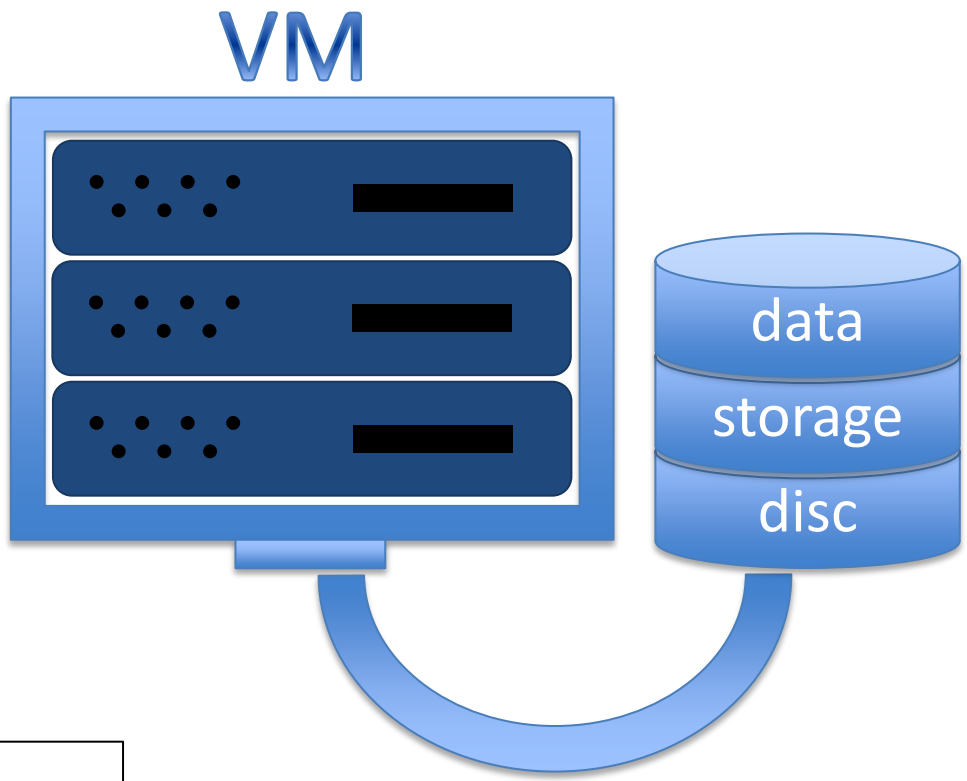
/dev/vda
Root disc
5-30 GB

For the operating system and software applications

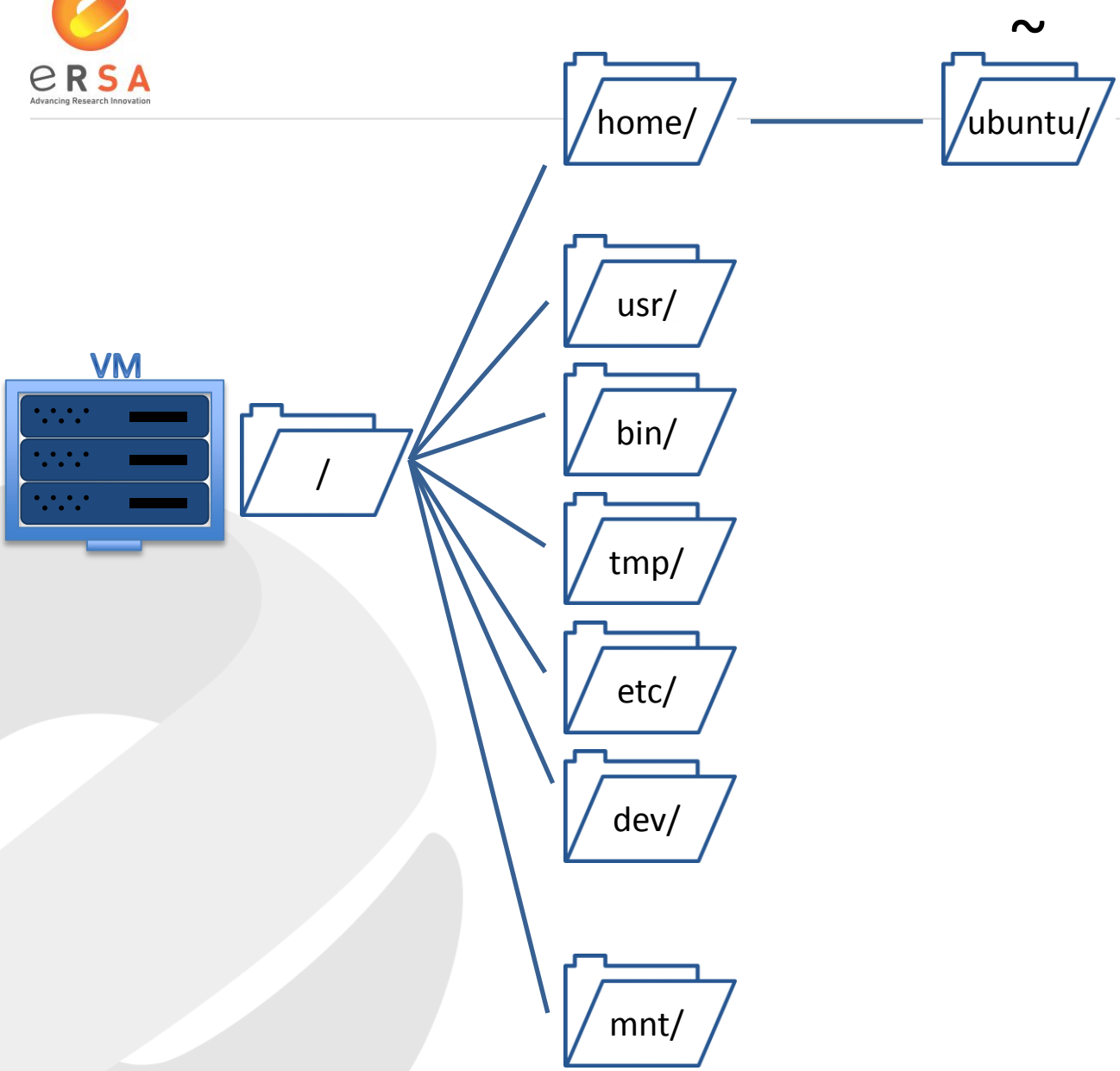


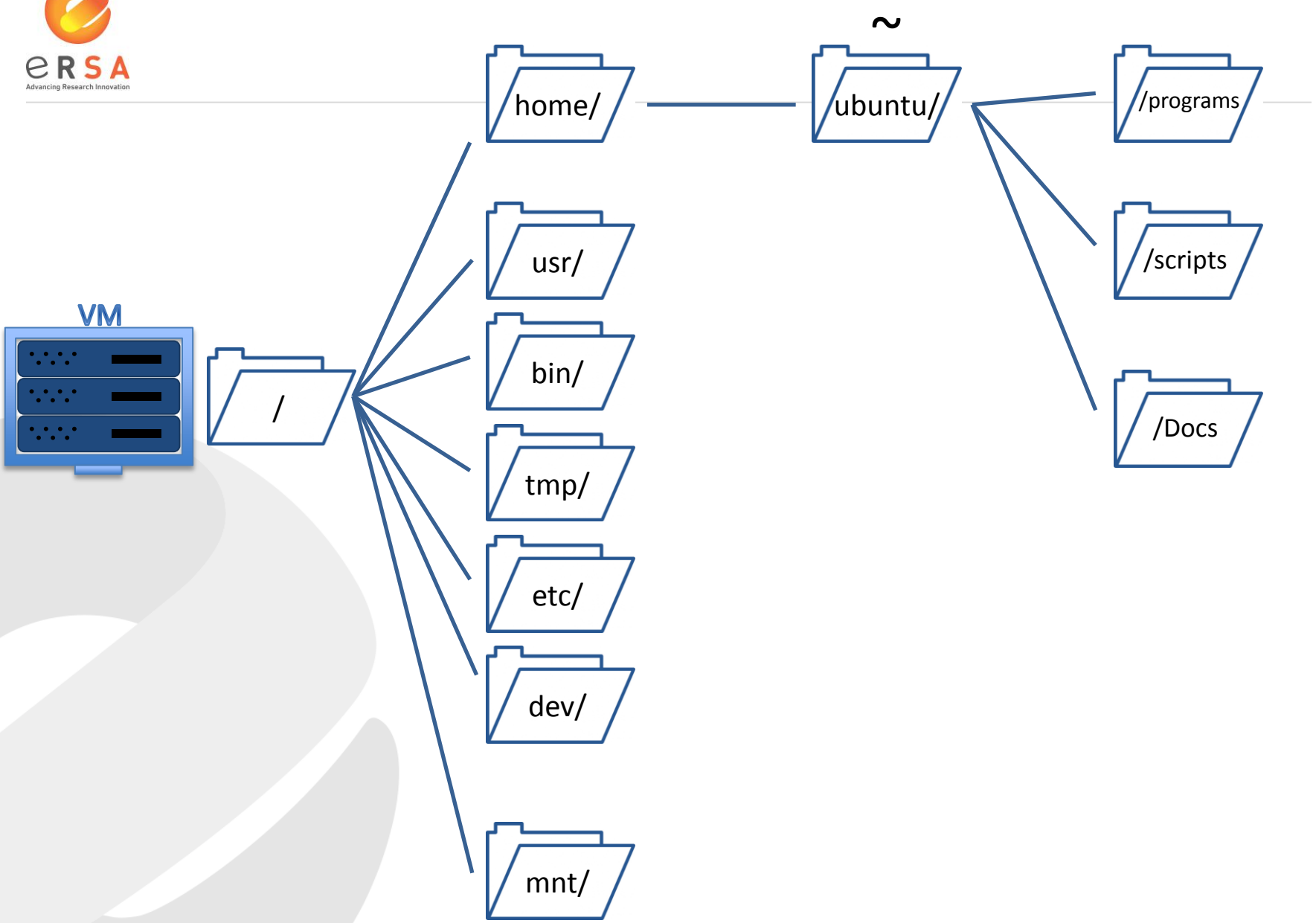
/dev/vda
Root disc
5-30 GB

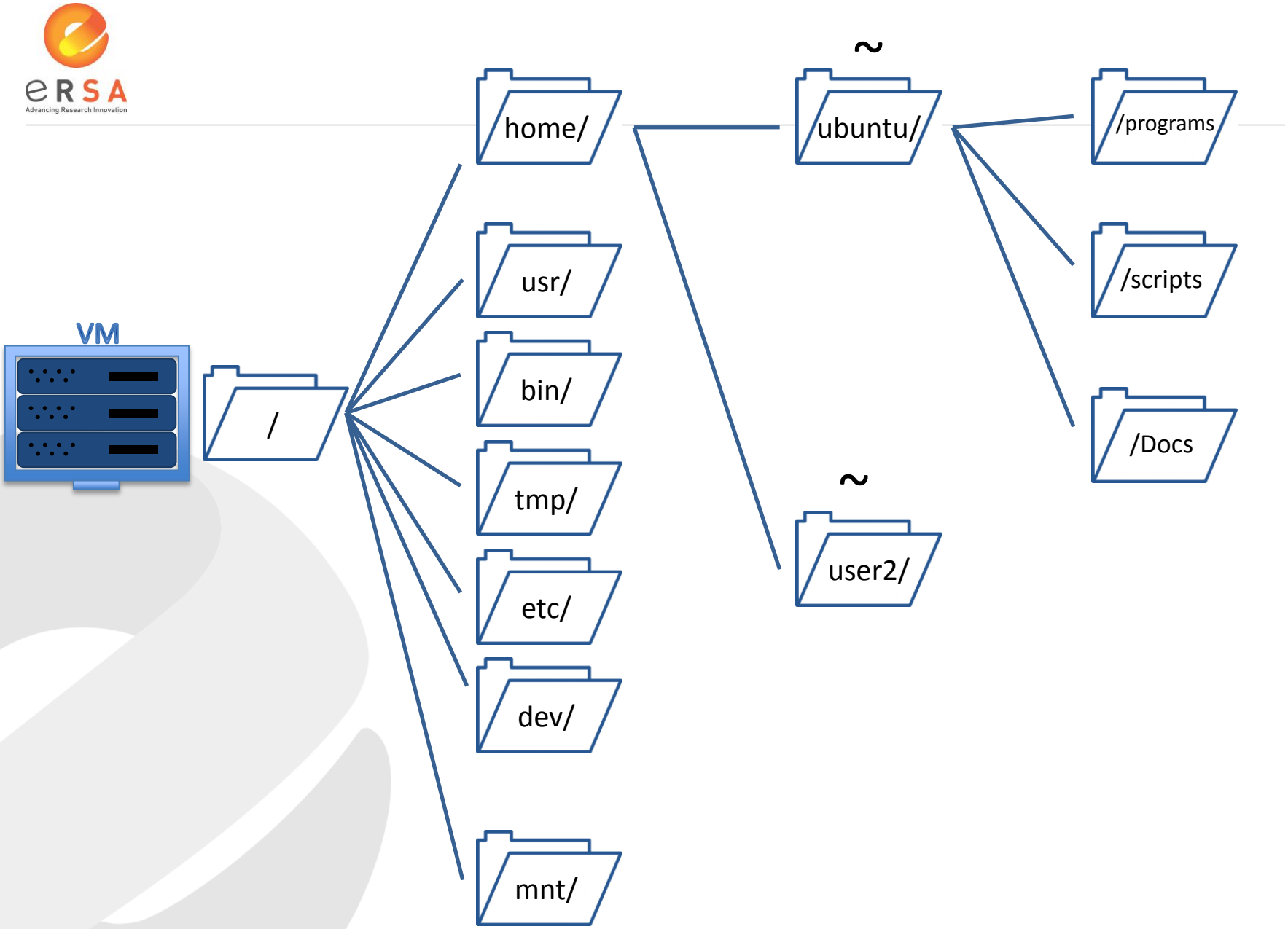
For the operating system and software applications

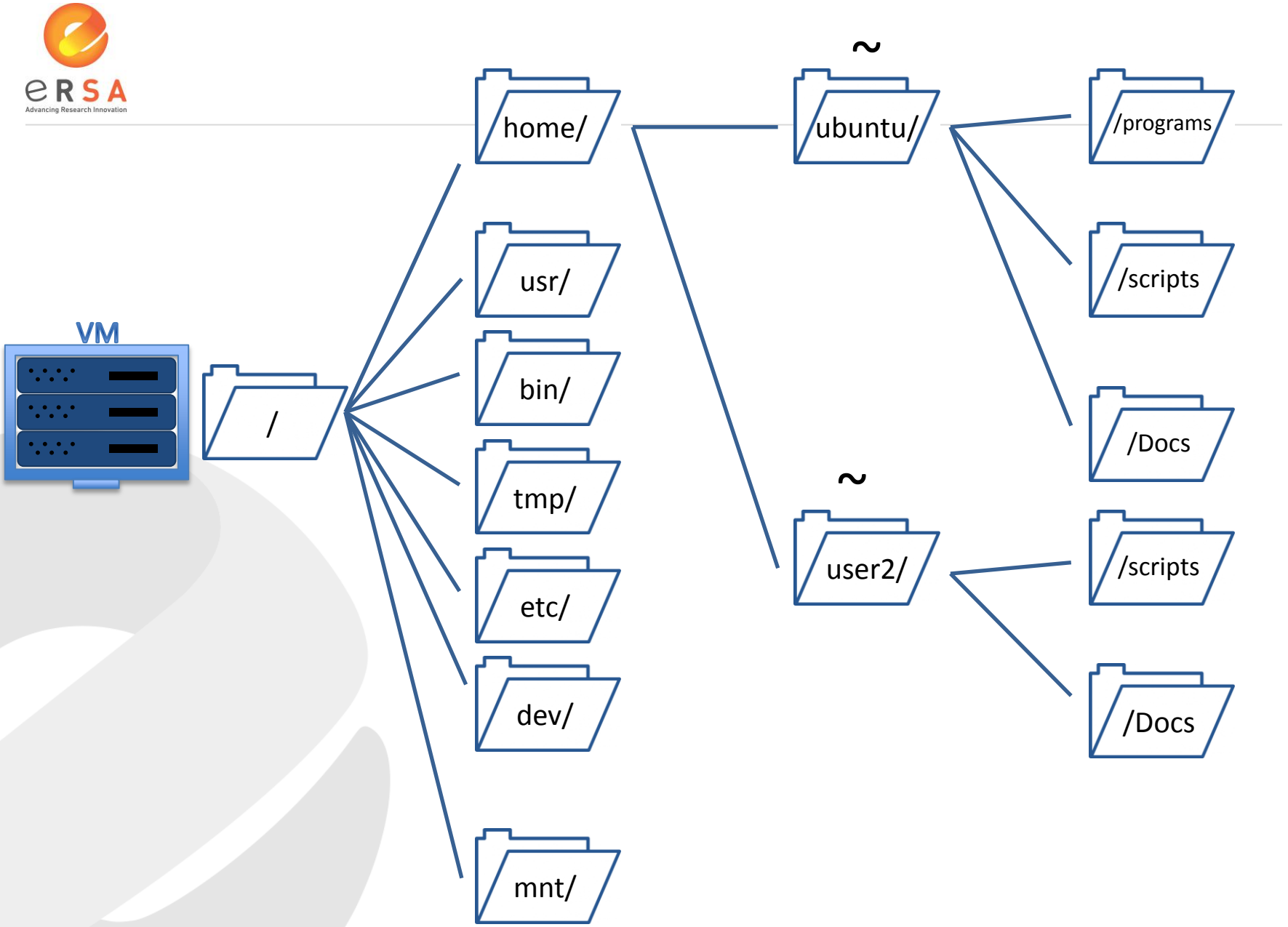


/
Root Directory
Includes user directories ($HOME directories)


~
home/
ubuntu/
VM
/
usr/
bin/
tmp/
etc/
dev/
mnt/

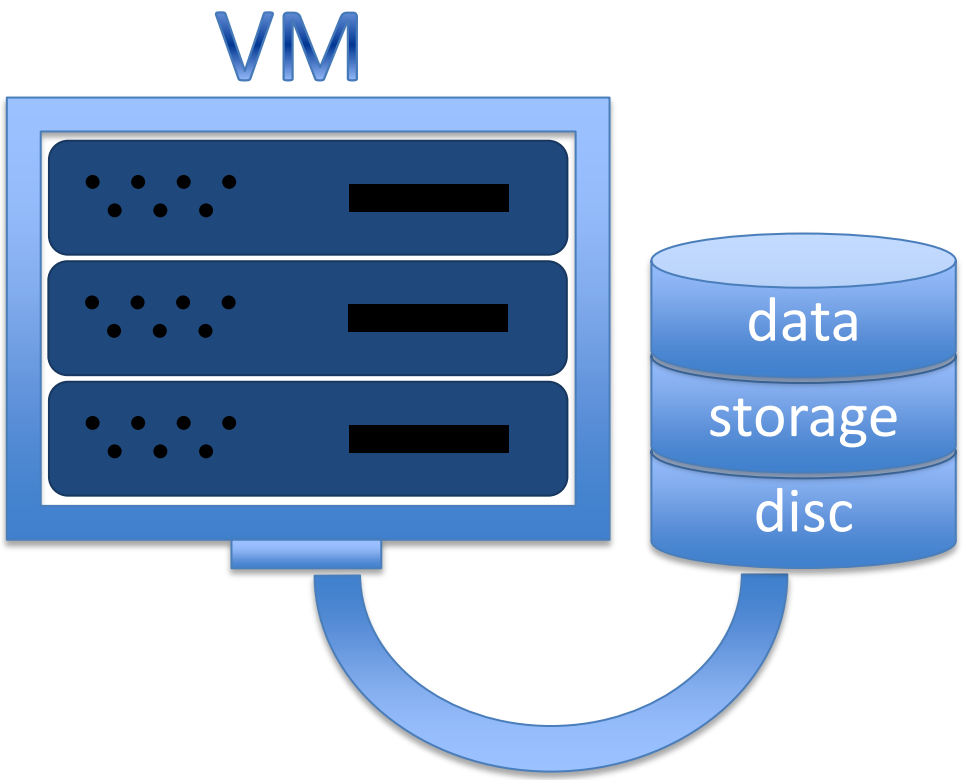

~
home/
ubuntu/
/programs
/scripts
/Docs
usr/
bin/
tmp/
etc/
dev/
mnt/
/
VM


eRSA
Advancing Research Innovation
VM
/
home/
usr/
bin/
tmp/
etc/
dev/
mnt/
~
ubuntu/
~
user2/
/programs
/scripts
/Docs

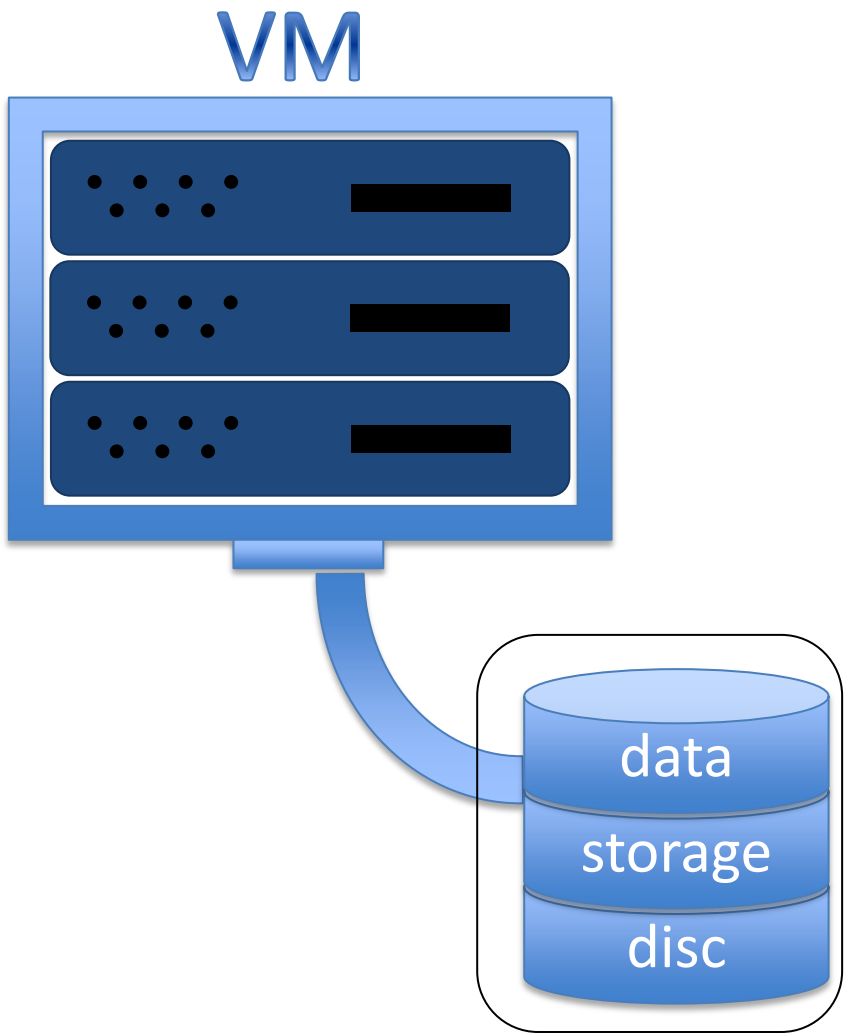

eRSA
Advancing Research Innovation
VM
/
home/
usr/
bin/
tmp/
etc/
dev/
mnt/
~
ubuntu/
/programs
/scripts
/Docs
~
user2/
/scripts
/Docs

/dev/vda
Root disc
5-30 GB

For the operating system and software applications

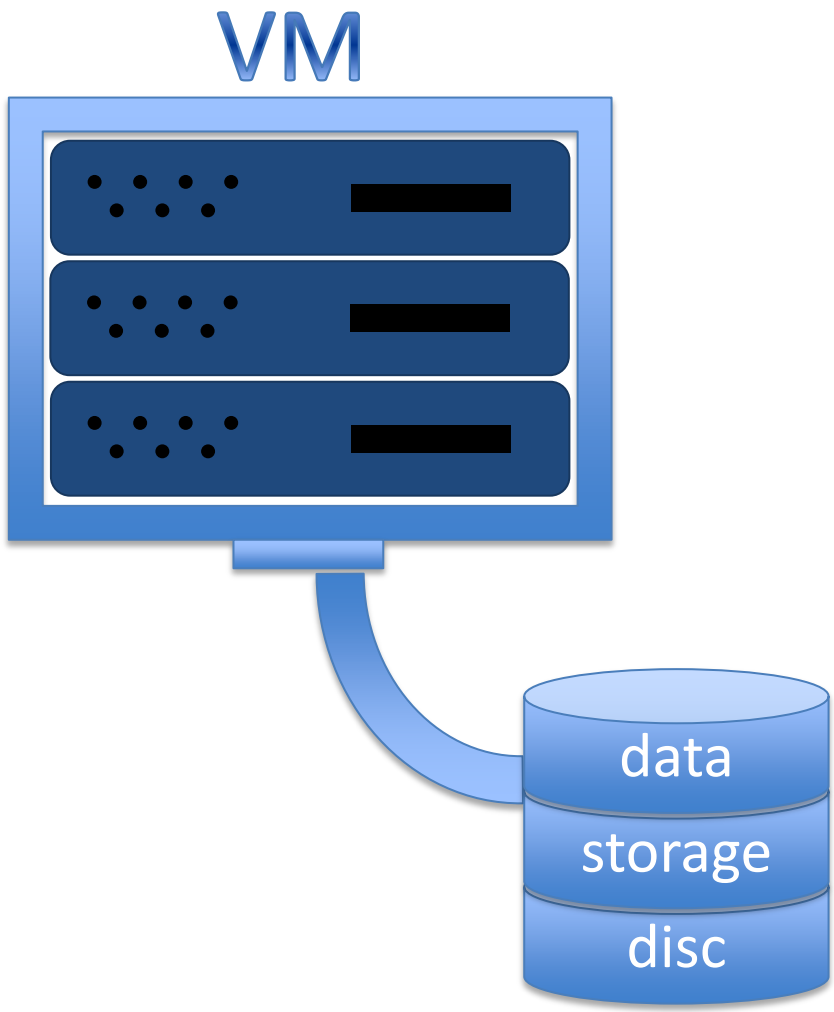


/dev/vda
Root (primary) disc
5-30 GB



/dev/vdb
Secondary storage
0 – 480 GB

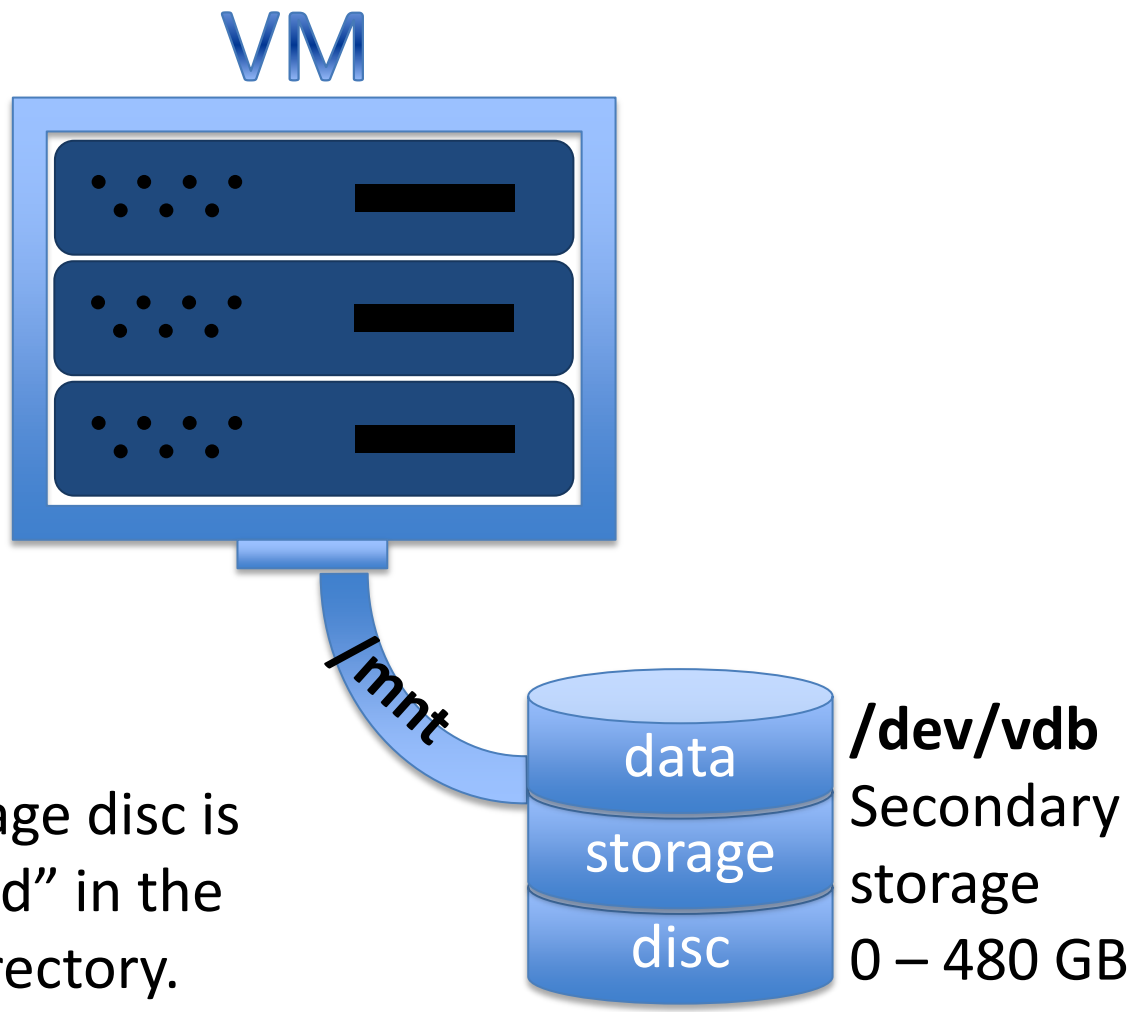
/dev/vda
Root (primary) disc
5-30 GB



/dev/vdb
Secondary storage
0 – 480 GB

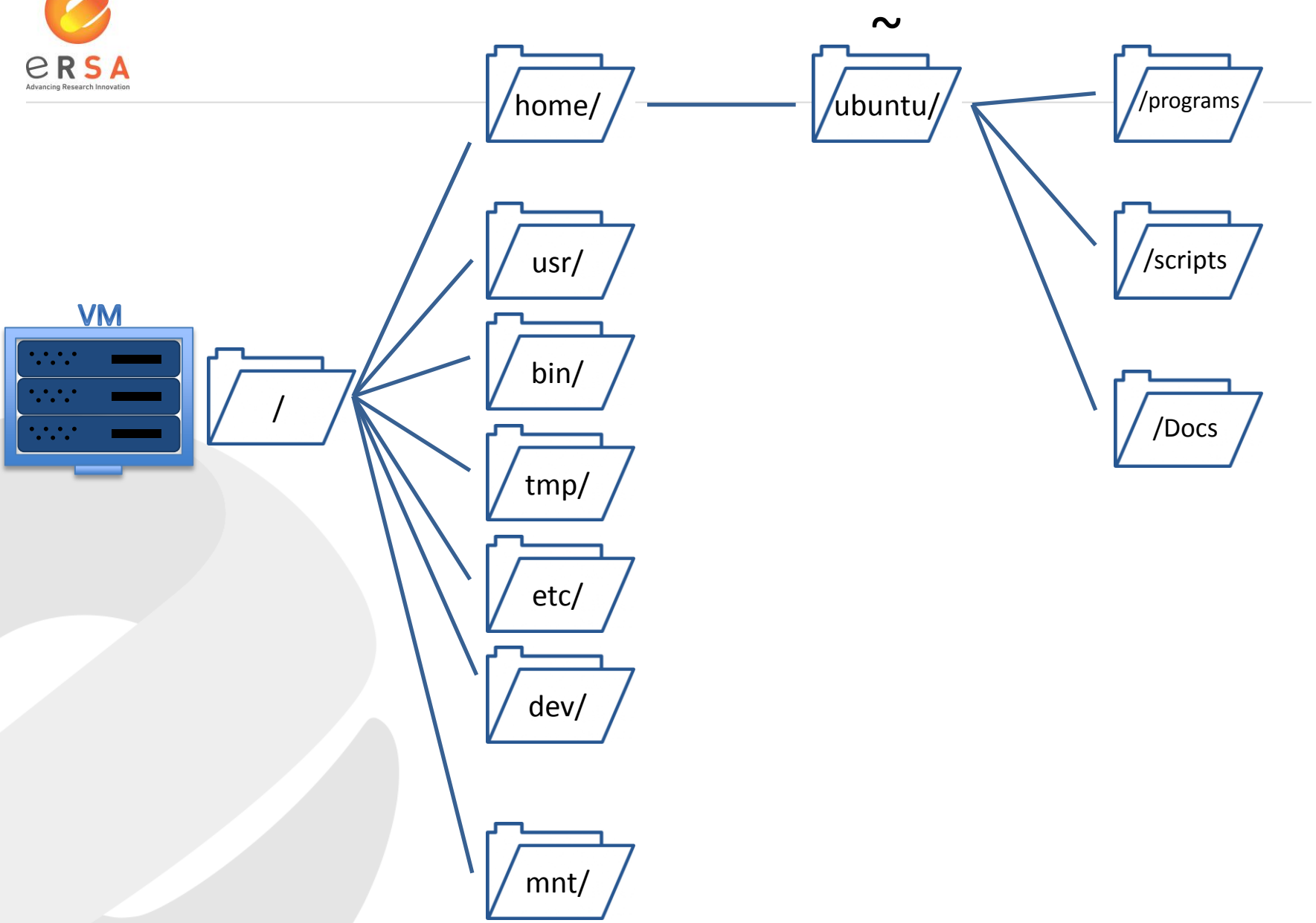
For storage of your data. Input and Output files.

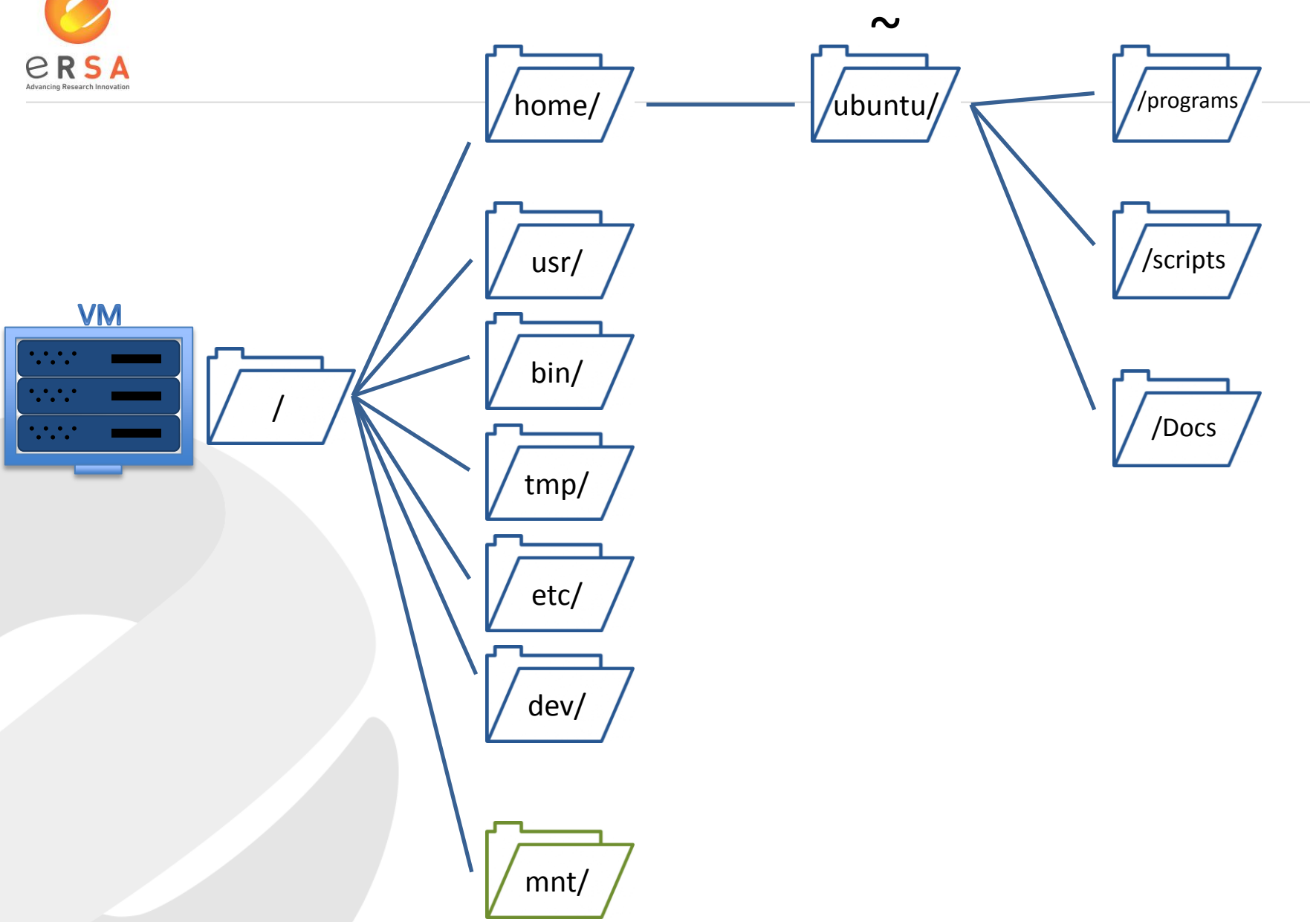
/dev/vda
Root (primary) disc
5-30 GB

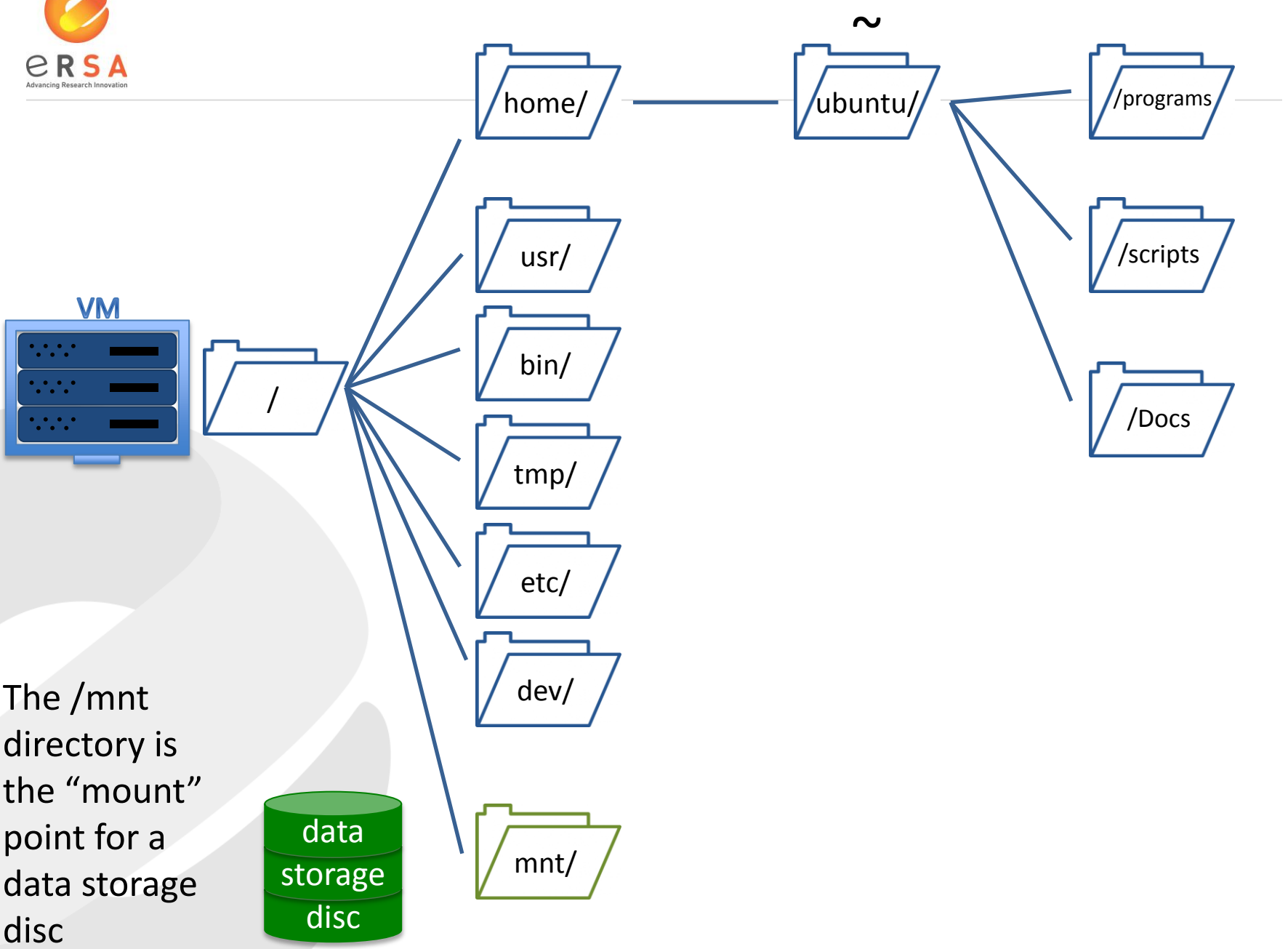


/dev/vdb
Secondary storage
0 – 480 GB

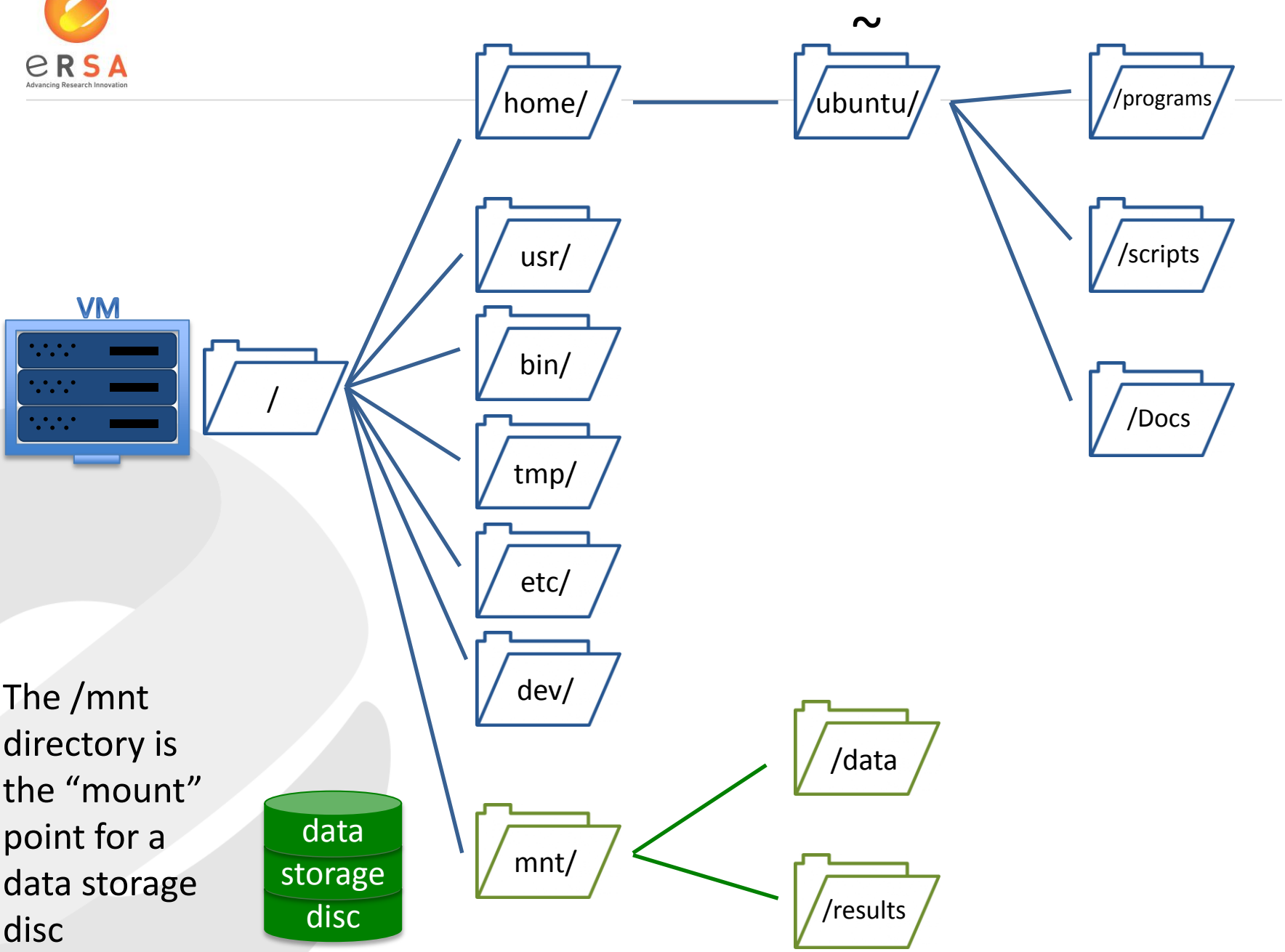
The storage disc is “mounted” in the **/mnt** directory.


~
home/
ubuntu/
/programs
/scripts
/Docs
usr/
bin/
tmp/
etc/
dev/
mnt/
/
VM


~
home/
ubuntu/
/programs
/scripts
/Docs
usr/
bin/
tmp/
etc/
dev/
mnt/
VM
/

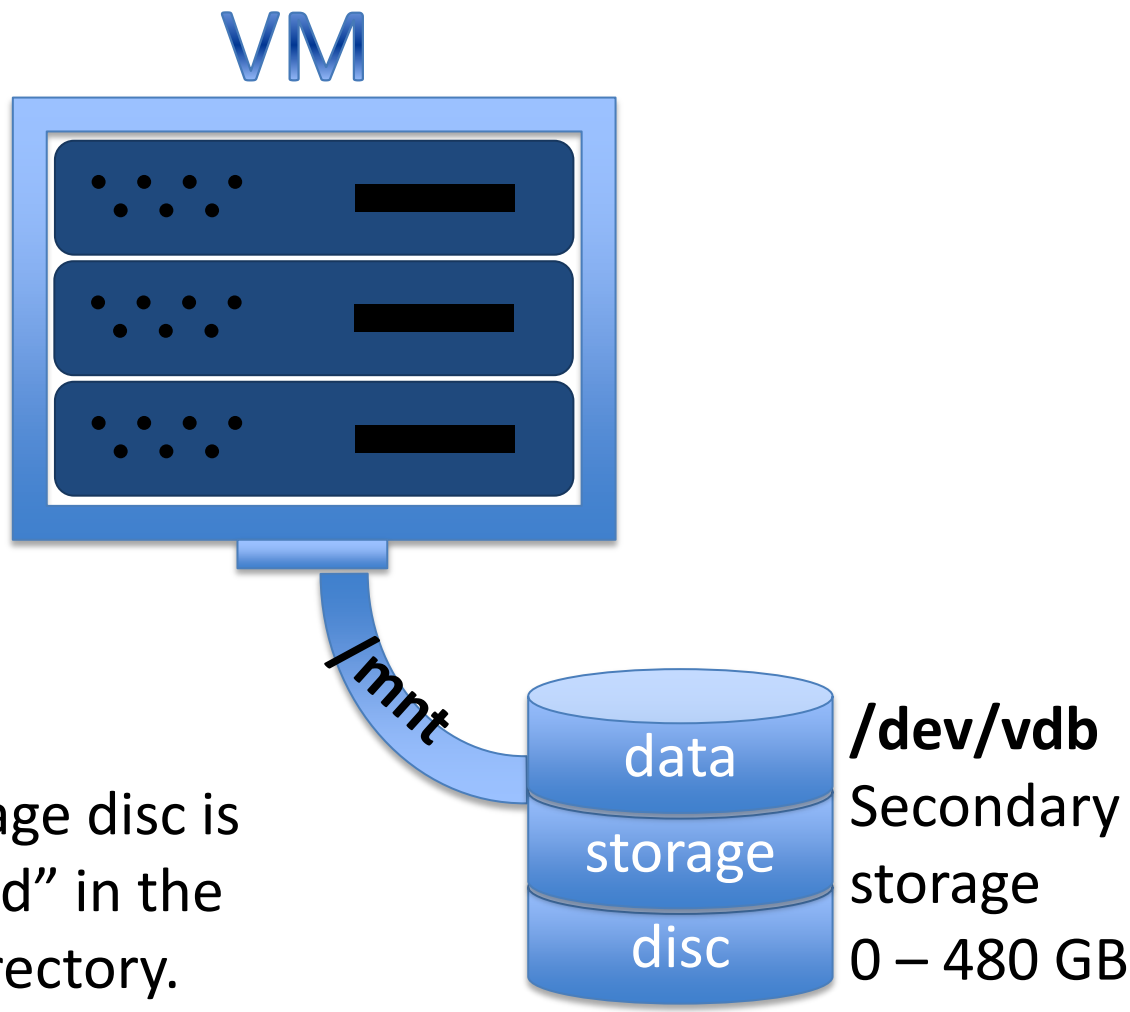


The /mnt directory is the “mount” point for a data storage disc



The /mnt directory is the “mount” point for a data storage disc

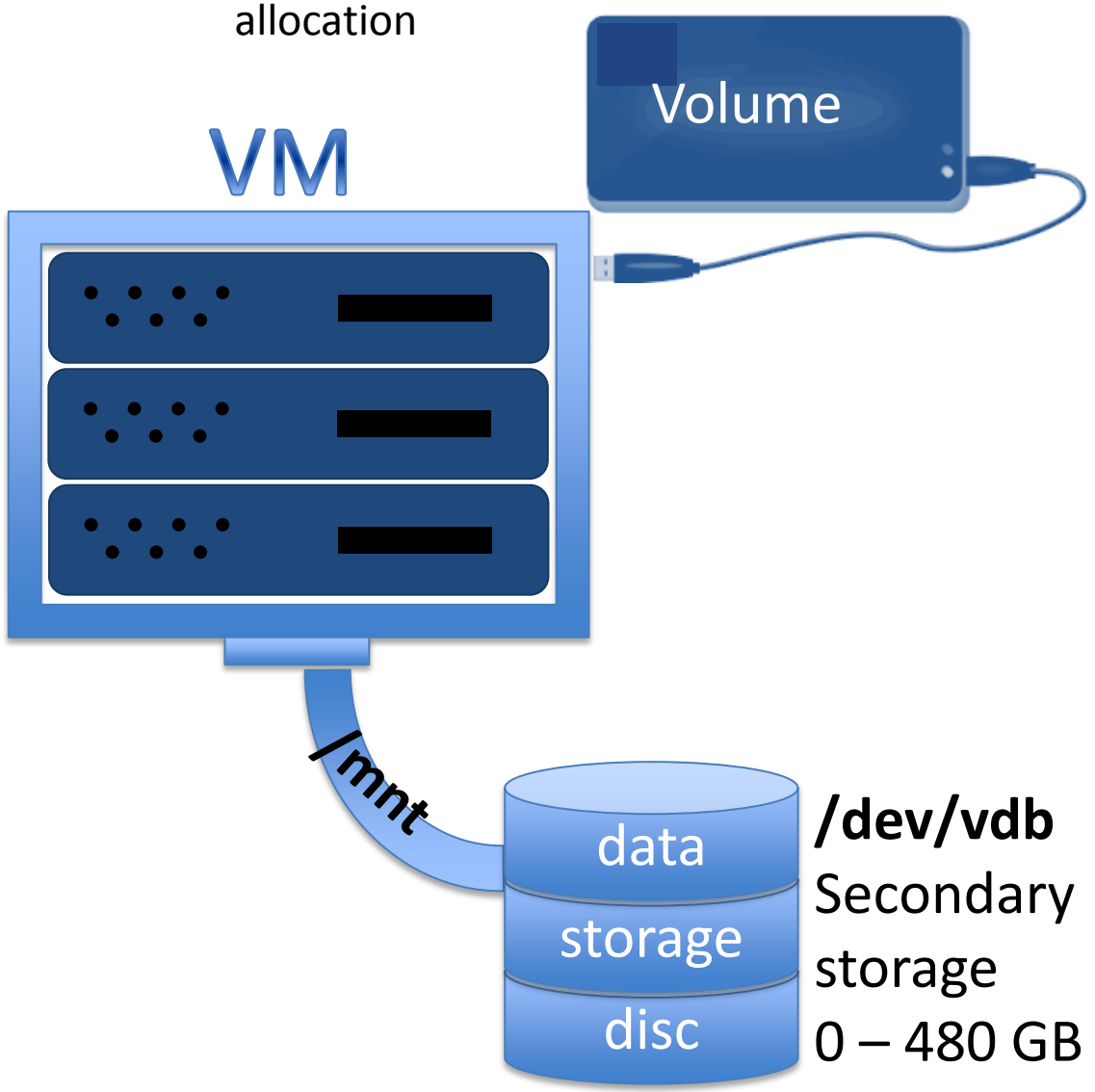
/dev/vda
Root (primary) disc
5-30 GB



/dev/vdb
Secondary storage
0 – 480 GB

The storage disc is “mounted” in the **/mnt** directory.

/dev/vdc
Volume storage
- available by allocation

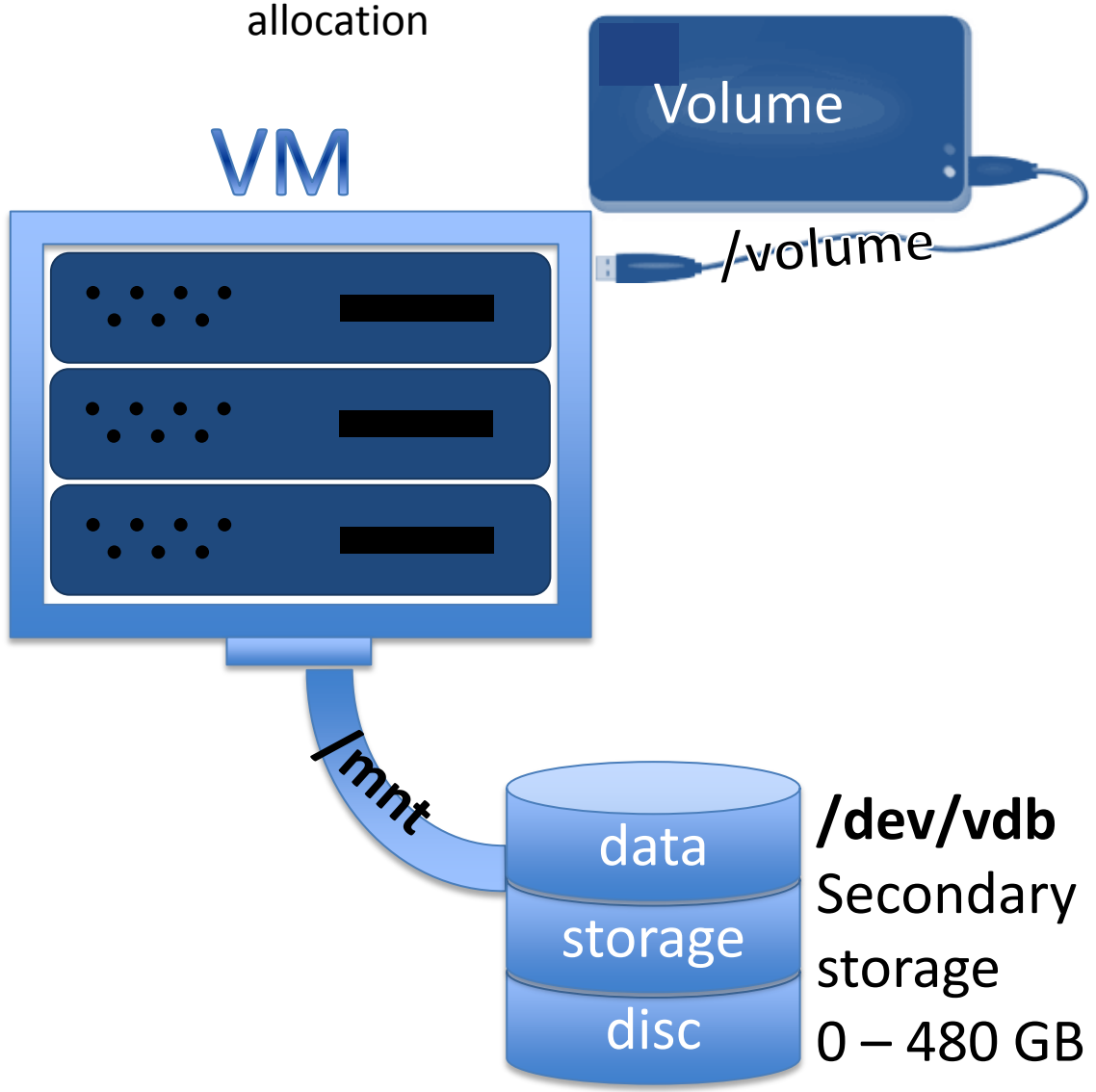


/dev/vda
Root (primary) disc
5-30 GB

/dev/vdb
Secondary storage
0 – 480 GB

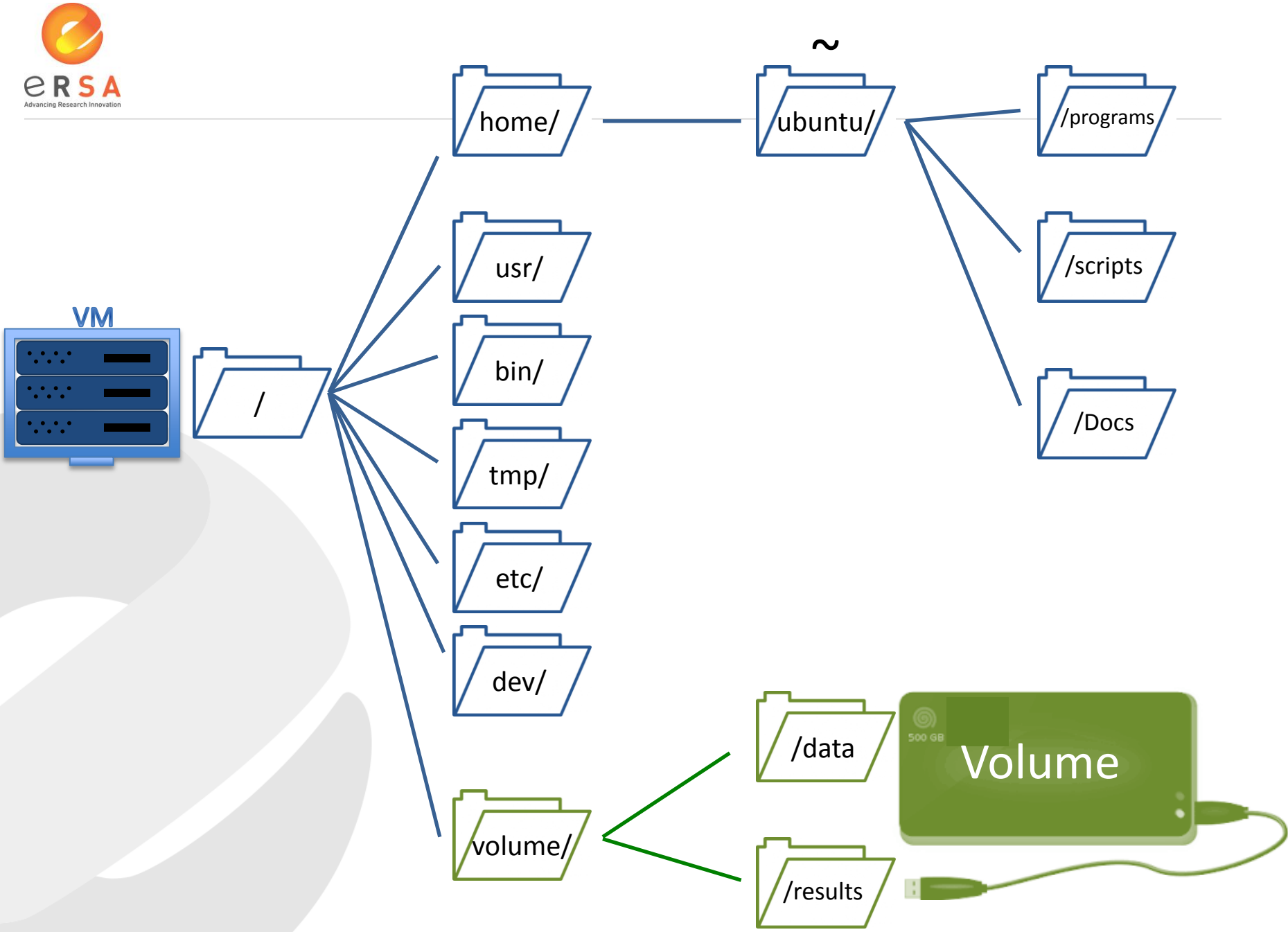
/dev/vdc
Volume storage
- available by allocation

Volume storage also must be mounted in a root directory



/dev/vda
Root (primary) disc
5-30 GB

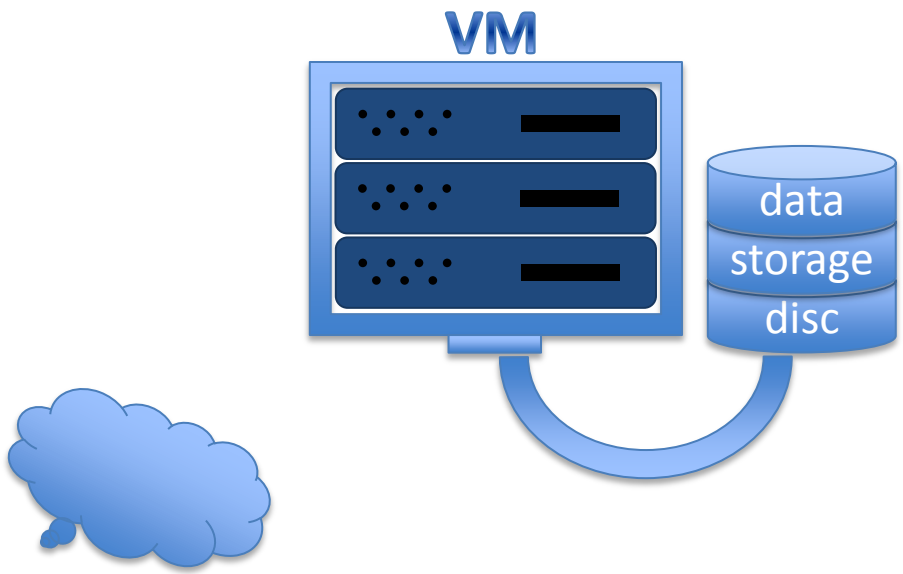
/dev/vdb
Secondary storage
0 – 480 GB

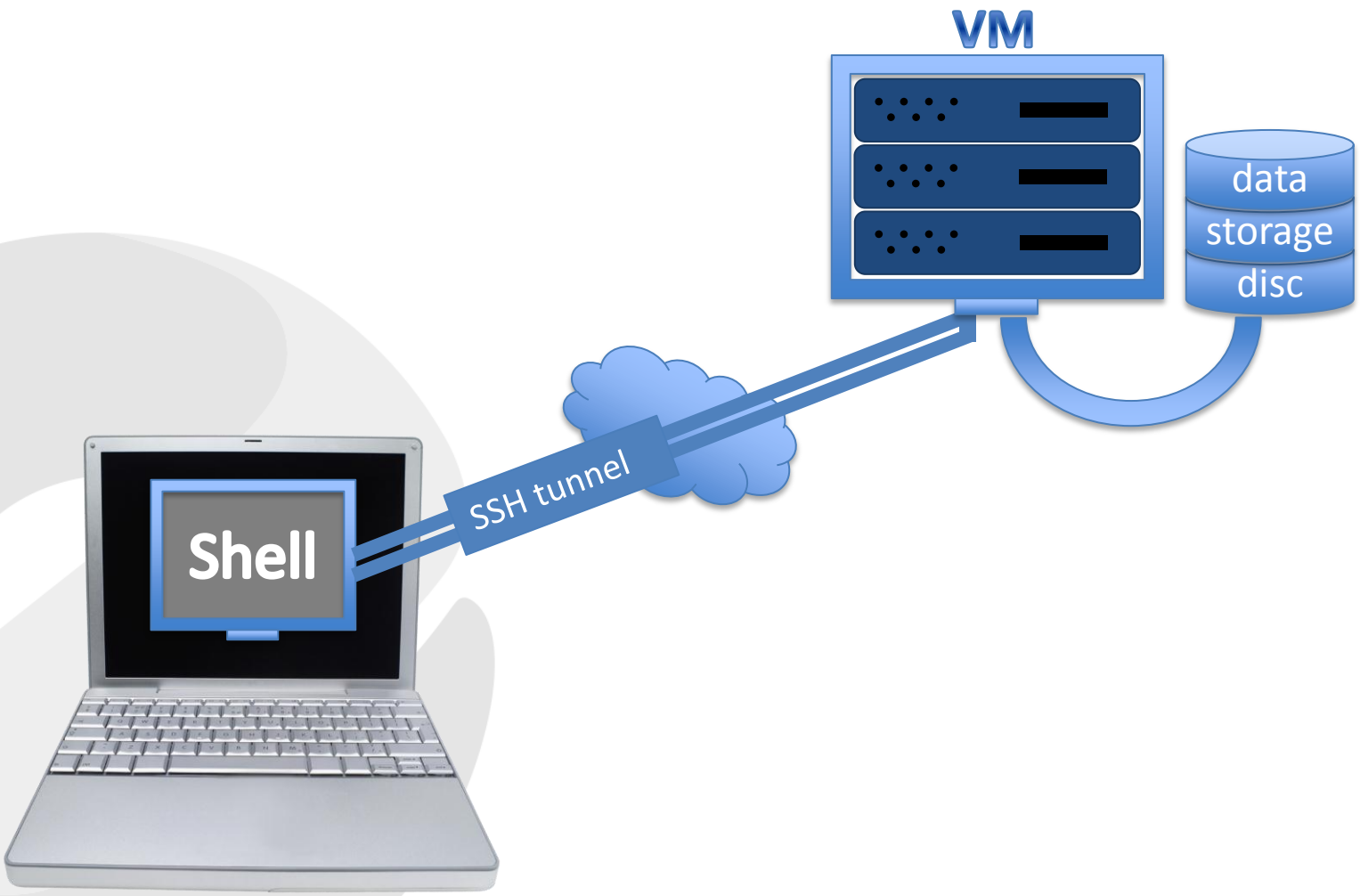

eRSA
Advancing Research Innovation
VM
/
home/
~
ubuntu/
/programs
/scripts
/Docs
usr/
bin/
tmp/
etc/
dev/
volume/
/data
/results
500 GB
Volume

eRSA
Advancing Research Innovation

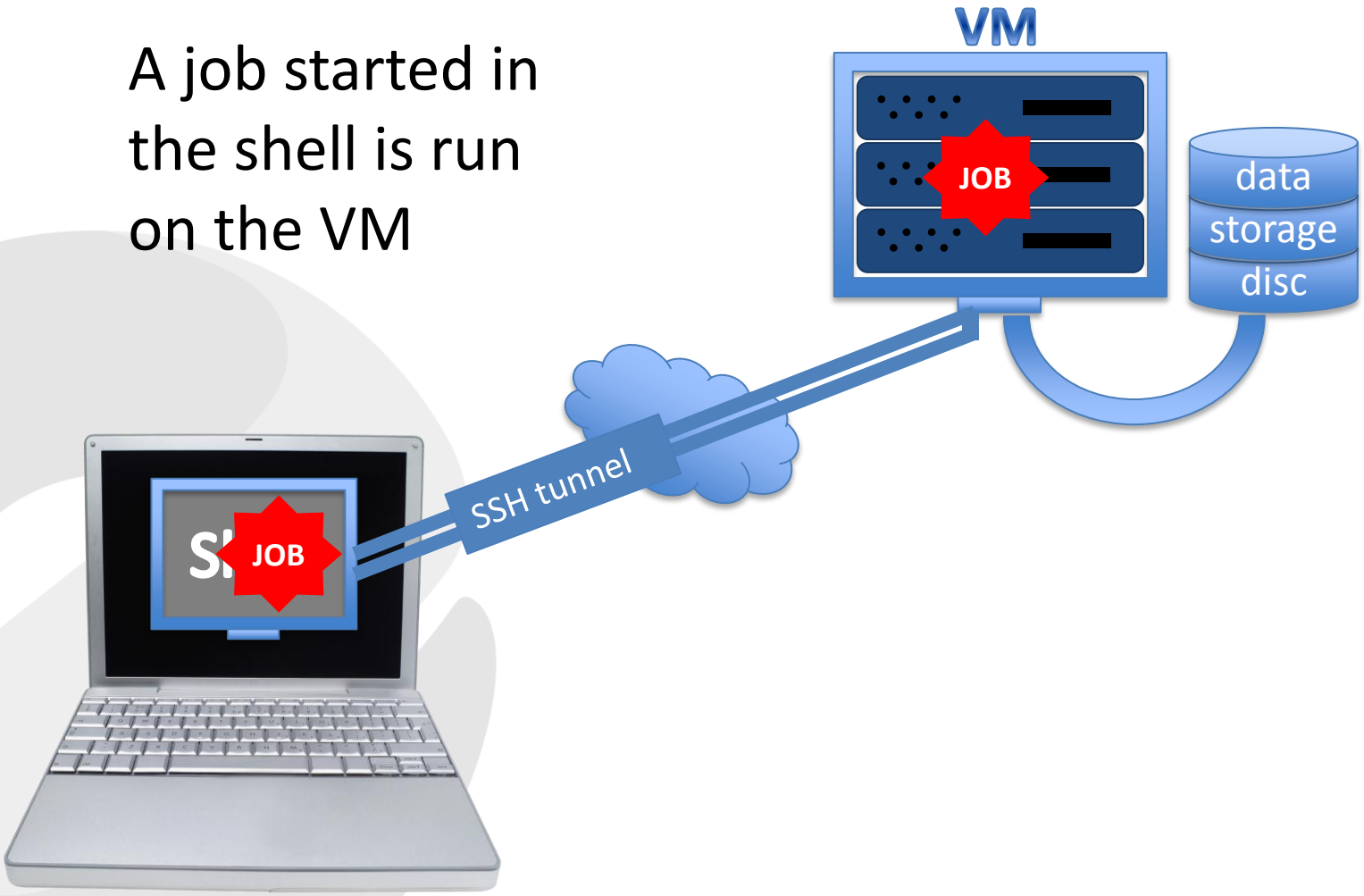
Running Jobs

Making sure your computing jobs keep running on your virtual machine when you are no longer connected to it


VM
data
storage
disc

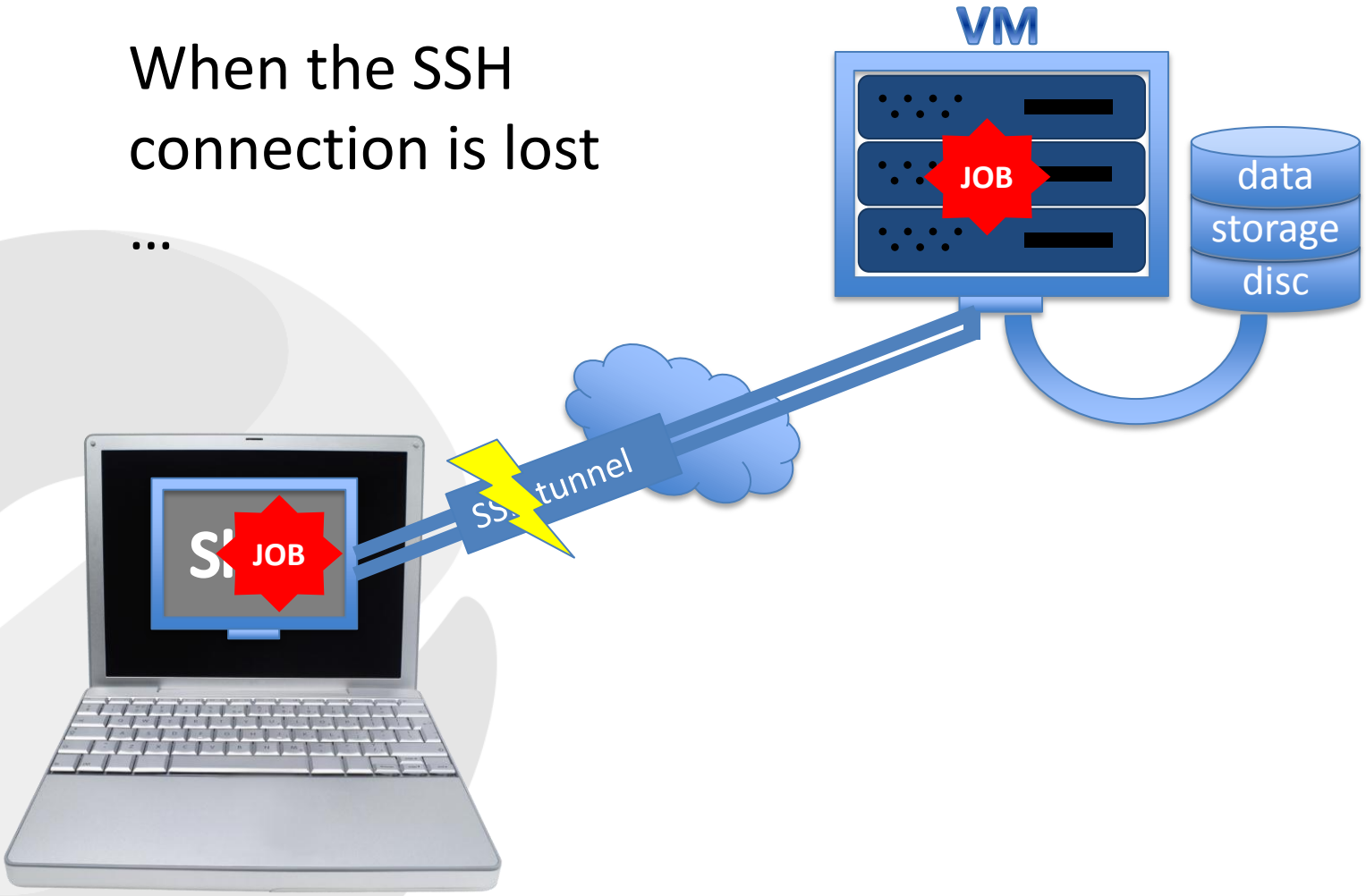

VM
data
storage
disc
SSH tunnel
Shell

A job started in the shell is run on the VM

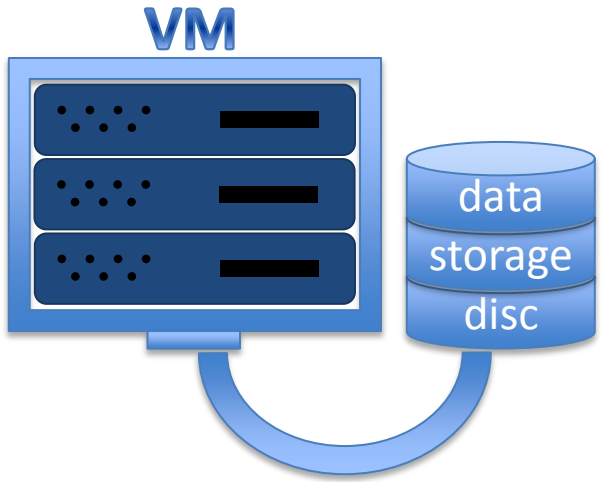


When the SSH connection is lost

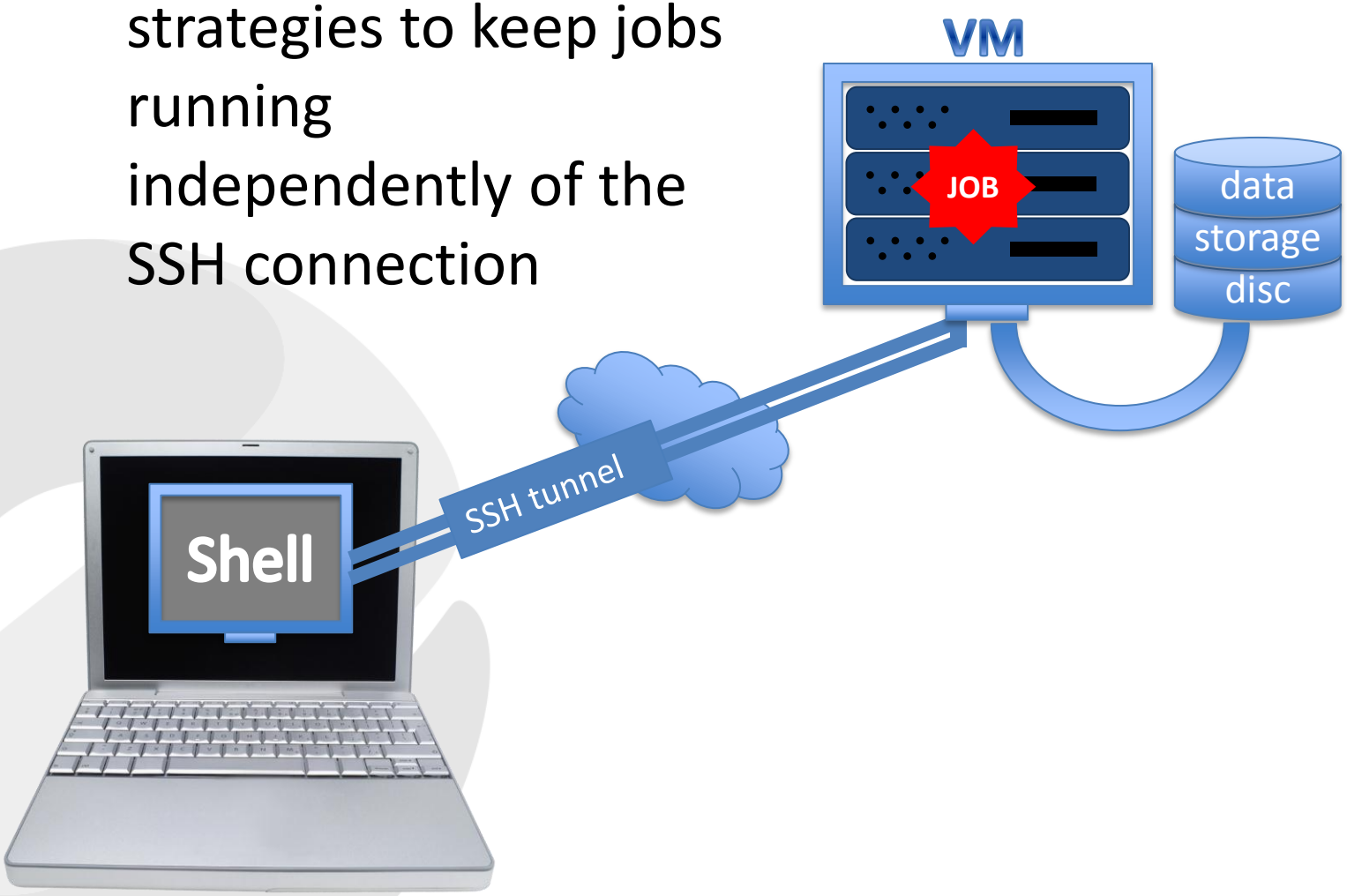
...



… the job will be discontinued



But there are strategies to keep jobs running independently of the SSH connection



eRSA
Advancing Research Innovation

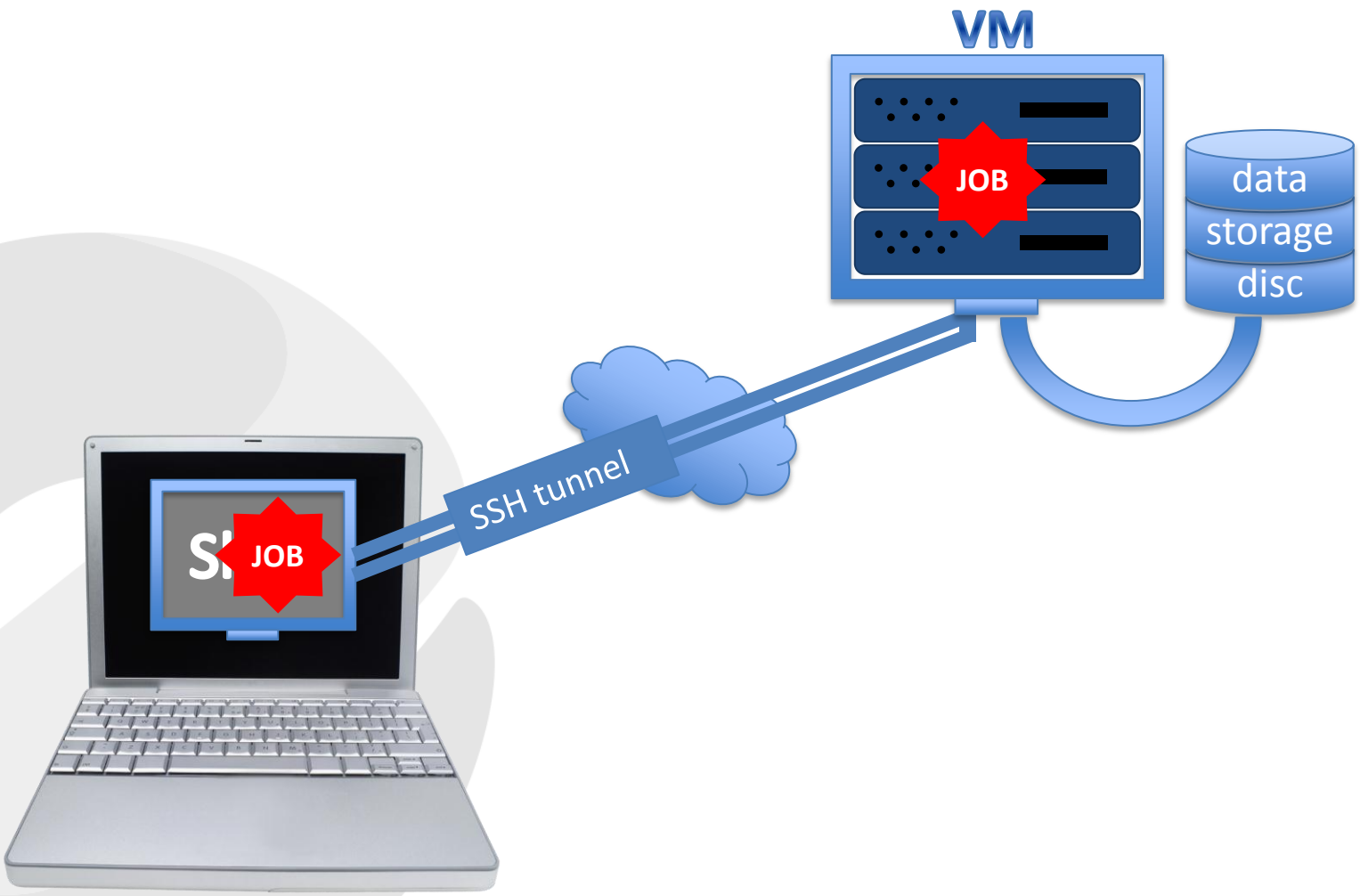
Running Jobs

1. Detach the job from the shell
   - Nohup, background, disown
2. Persistent virtual consoles
   - Screen, Tmux, Byobu

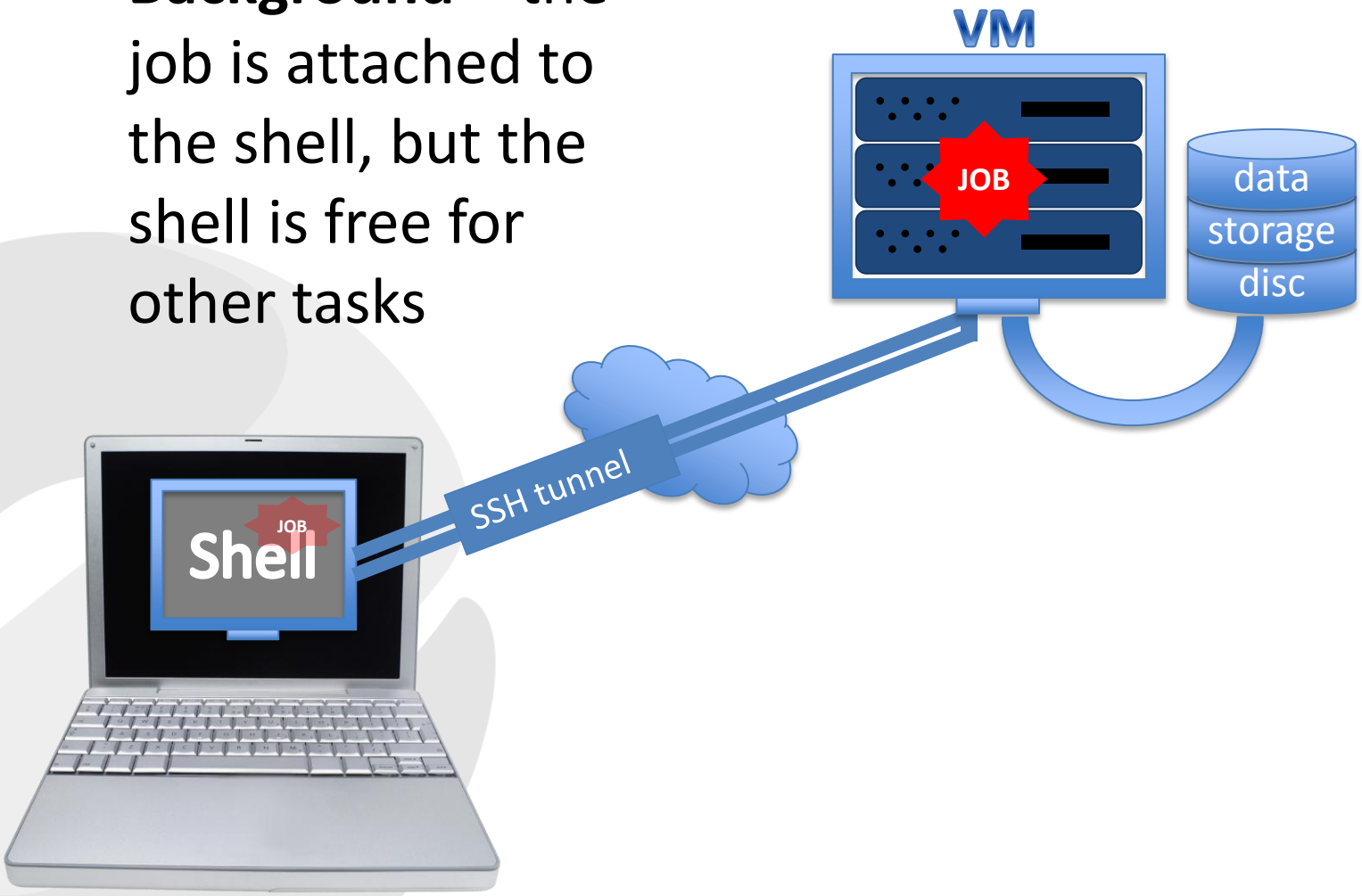
eRSA
Advancing Research Innovation

Detach the job from the shell

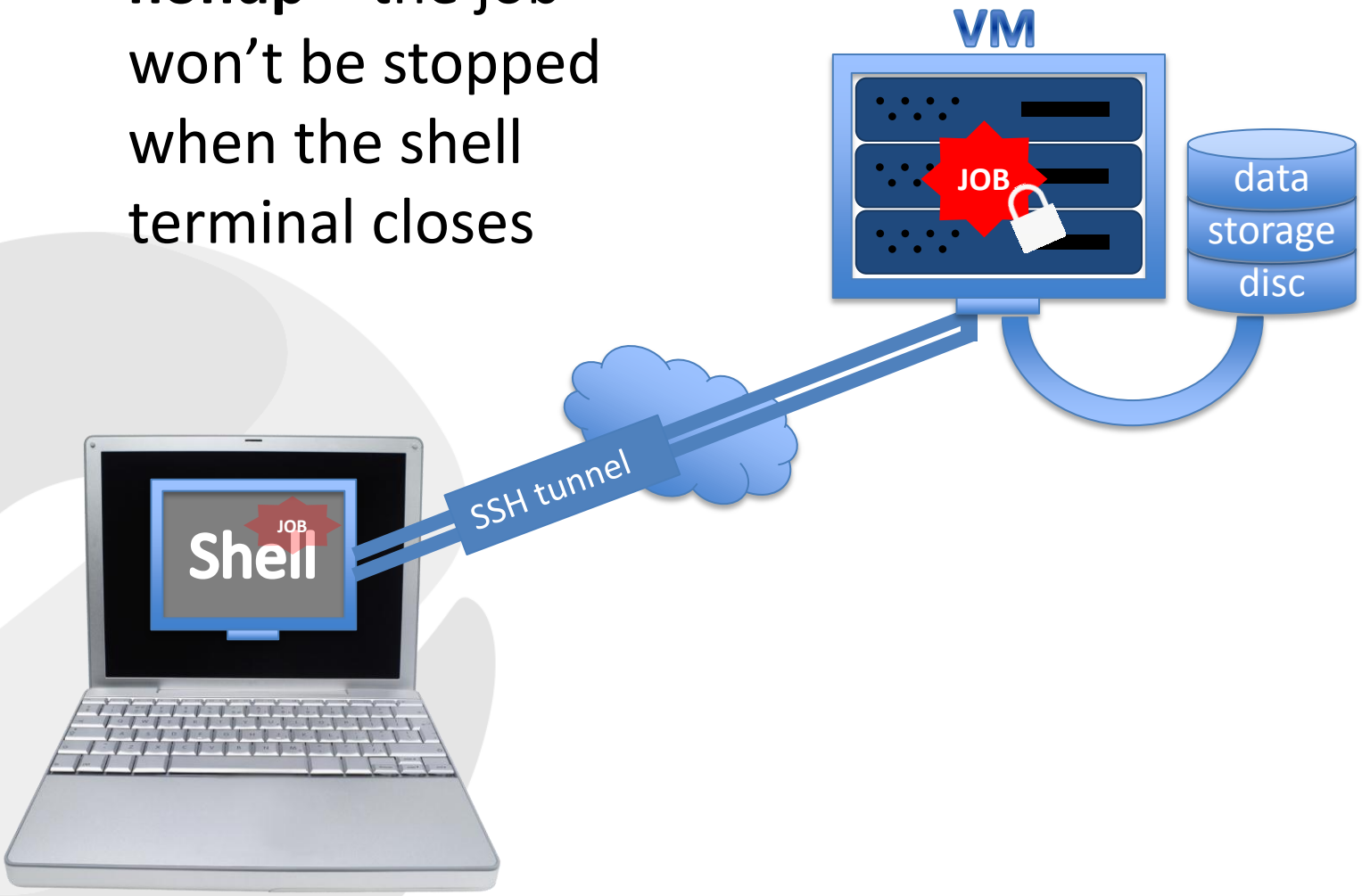
nohup


VM
JOB
data
storage
disc
SSH tunnel
JOB

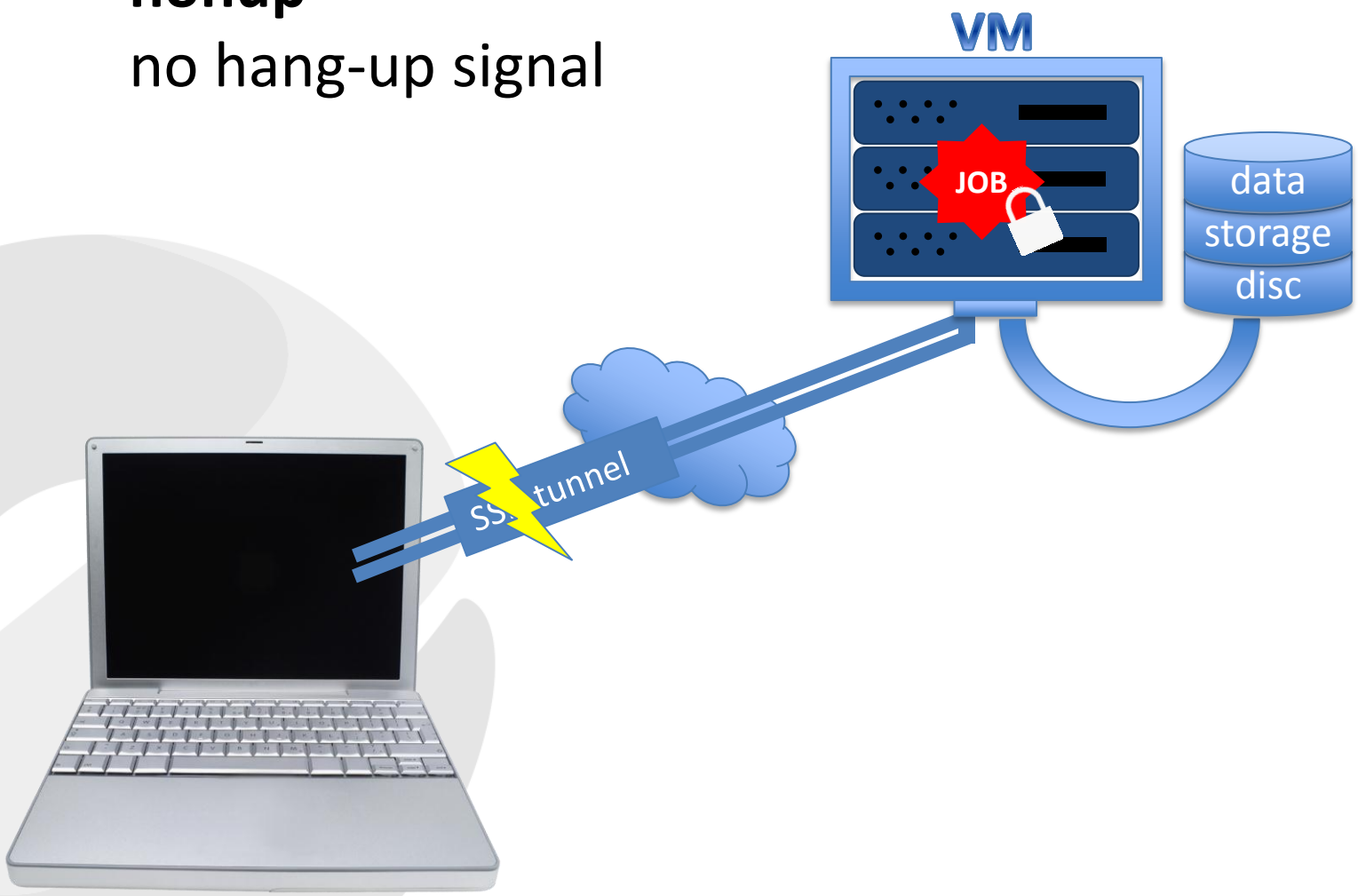
Background – the job is attached to the shell, but the shell is free for other tasks



nohup – the job won't be stopped when the shell terminal closes



nohup =
no hang-up signal



eRSA
Advancing Research Innovation

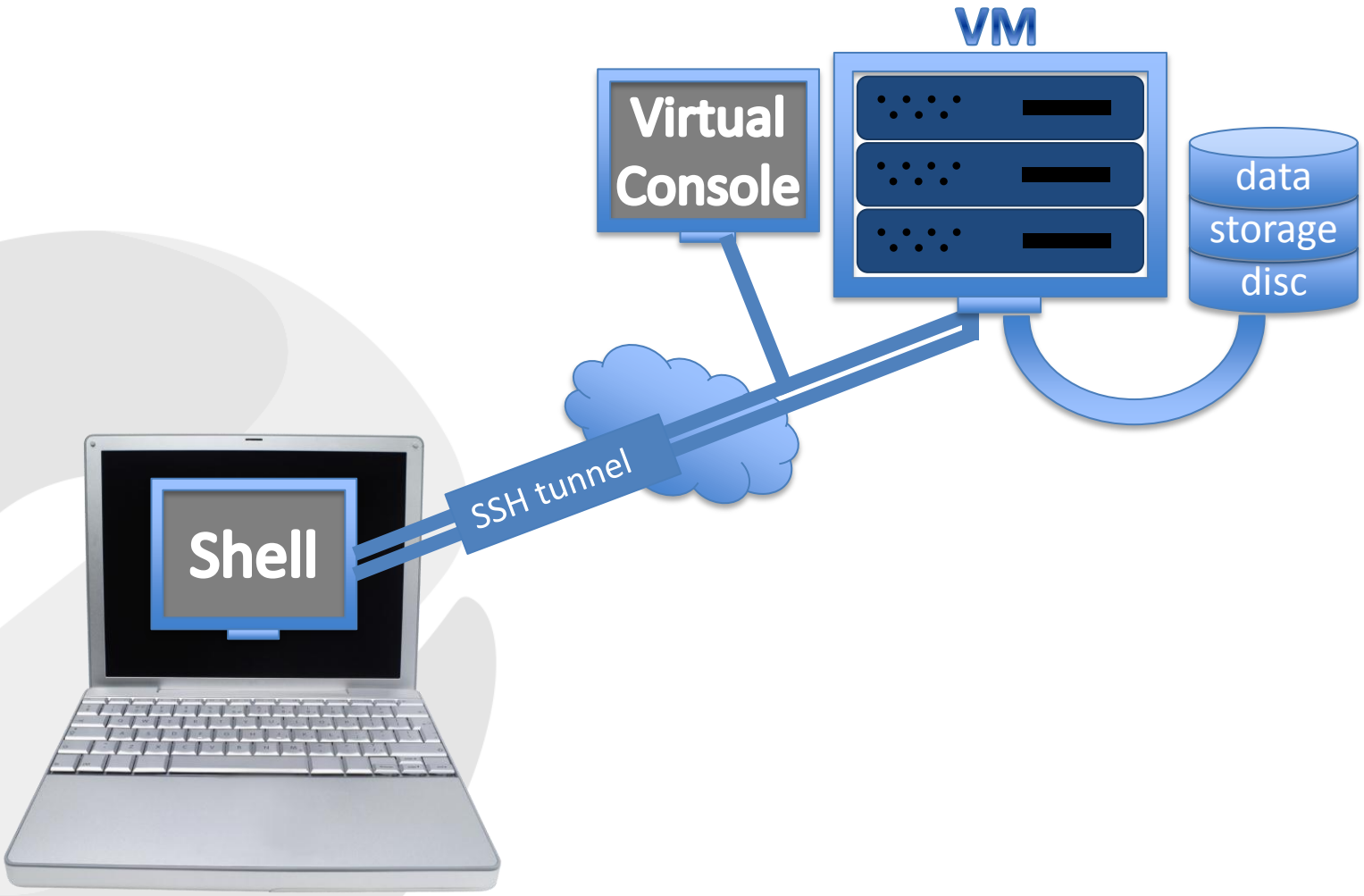
Persistent virtual consoles

GNU Screen

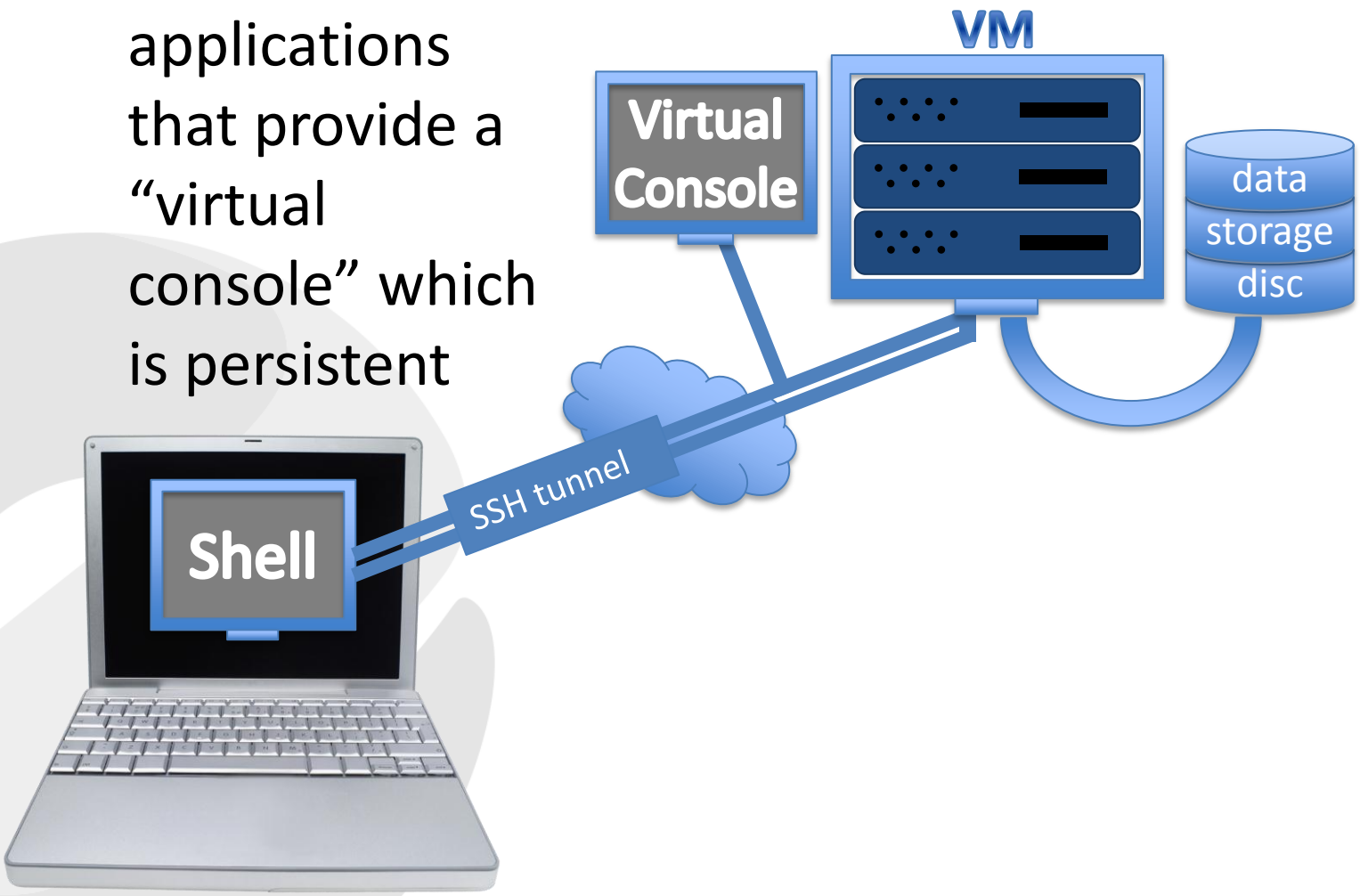
Tmux

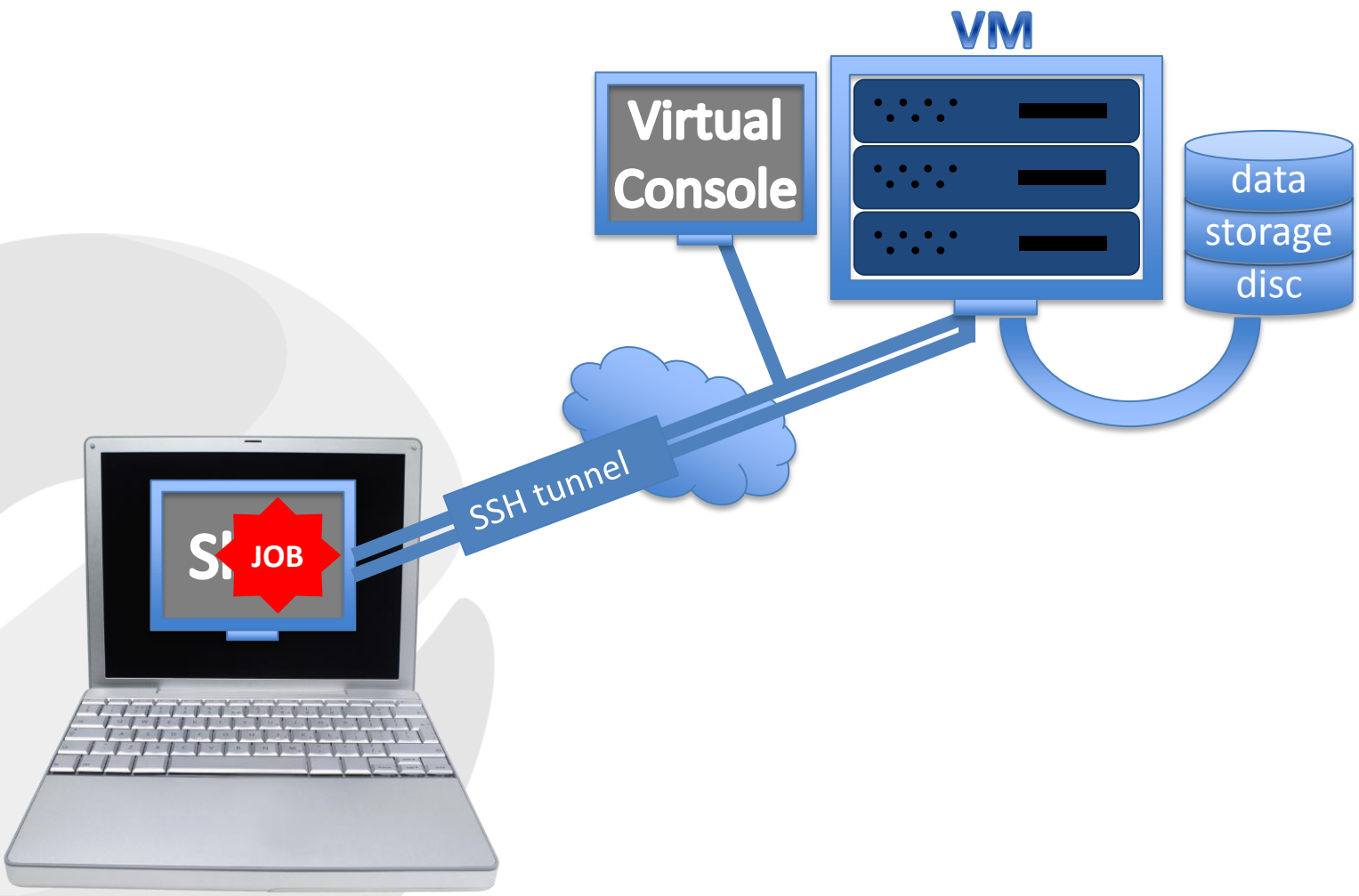
Byobu

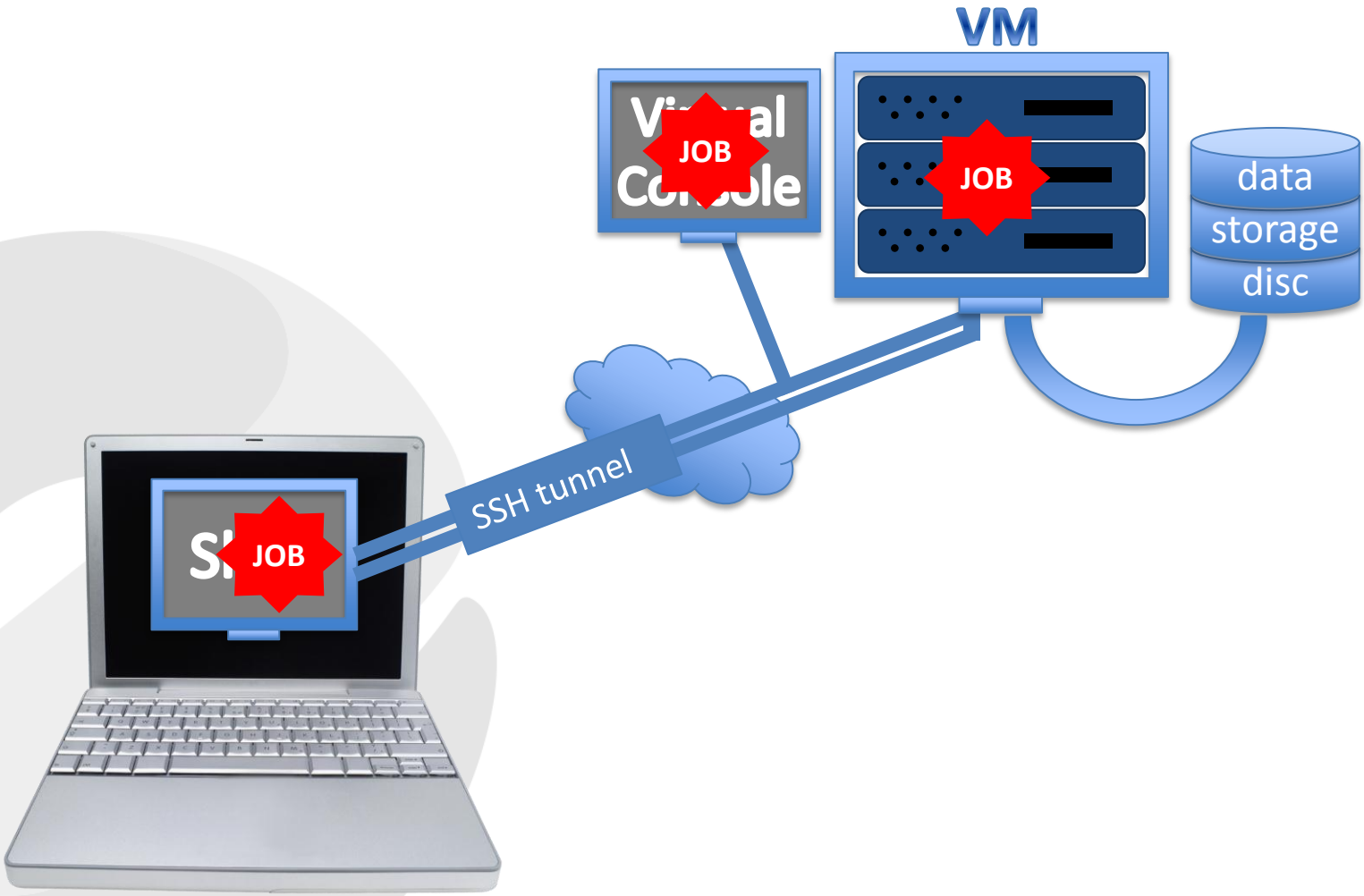
GNU Screen and Tmux

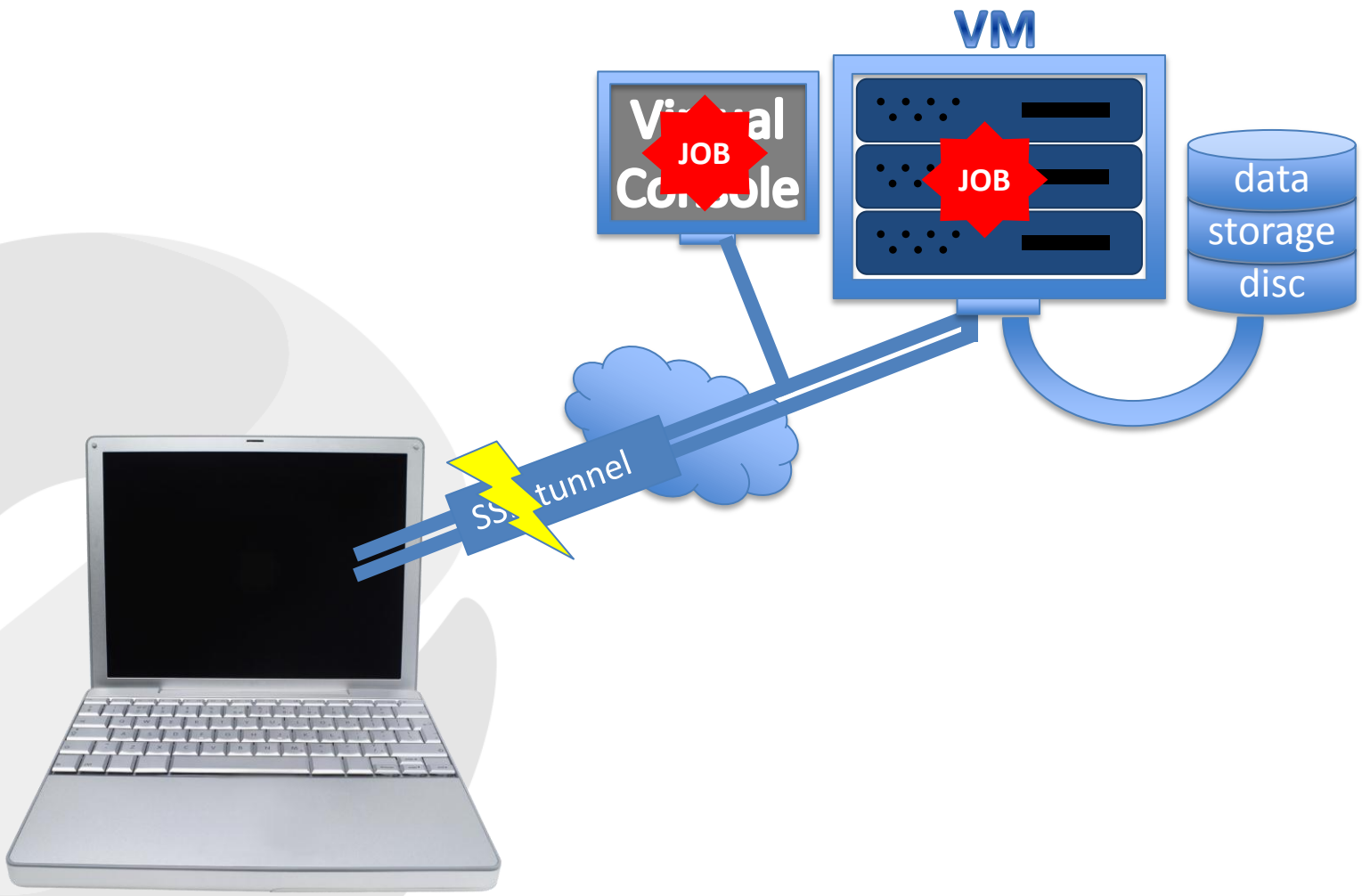


GNU Screen and **Tmux** – applications that provide a "virtual console" which is persistent

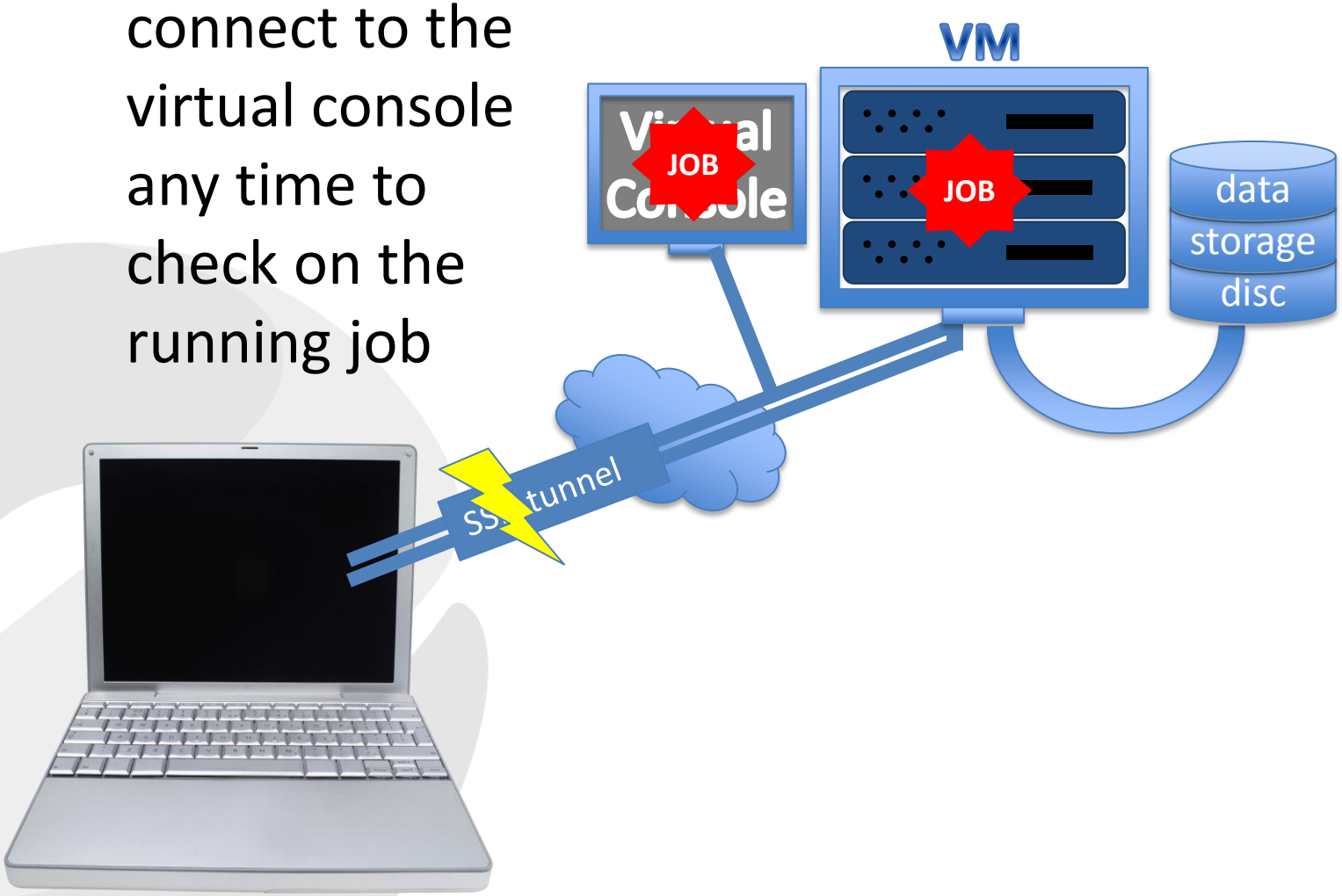



VM
Virtual Console
data
storage
disc
SSH tunnel
JOB

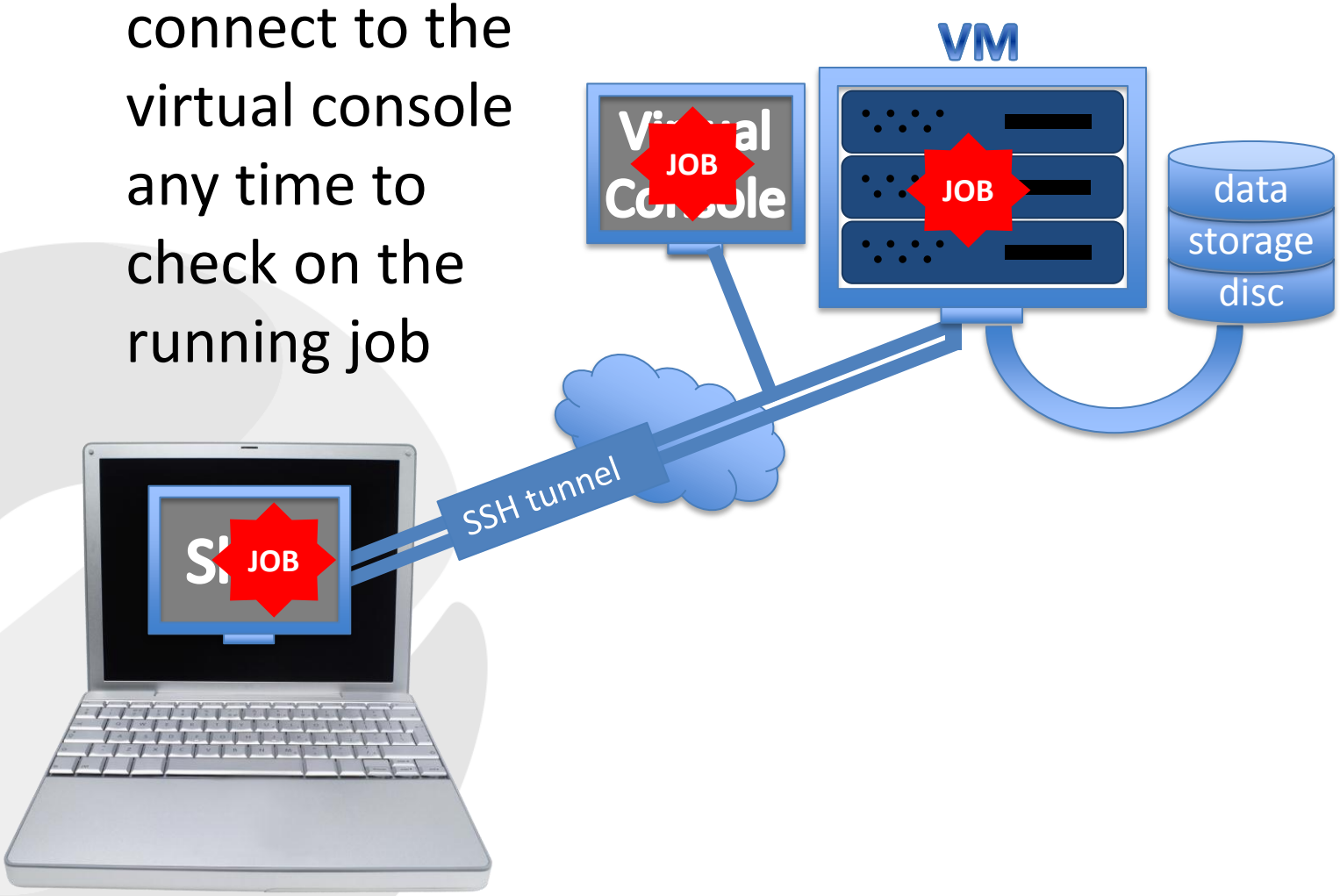

VM
JOB
JOB
data
storage
disc
SSH tunnel
JOB


VM
JOB
JOB
data
storage
disc
tunnel

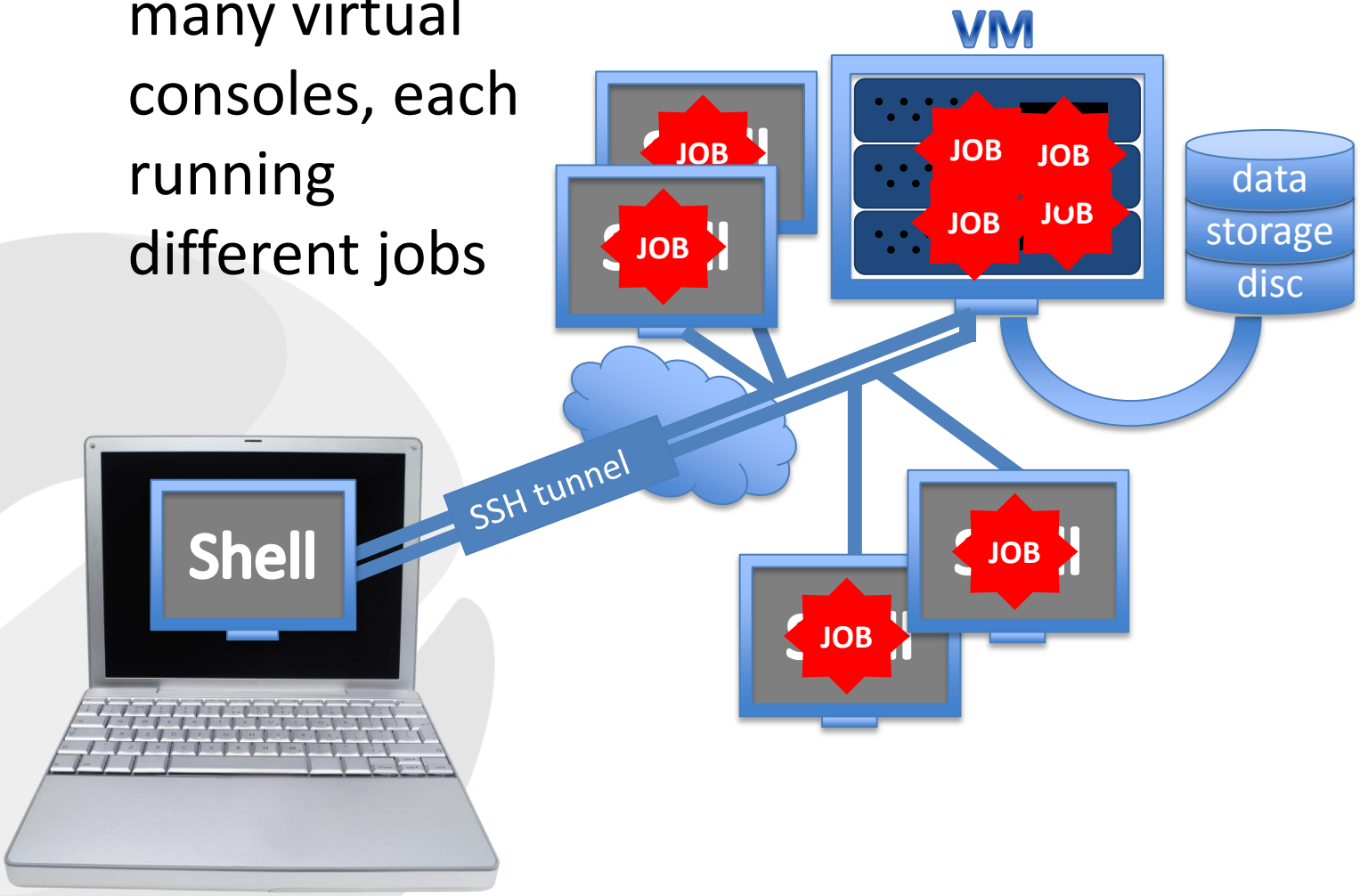
You can connect to the virtual console any time to check on the running job



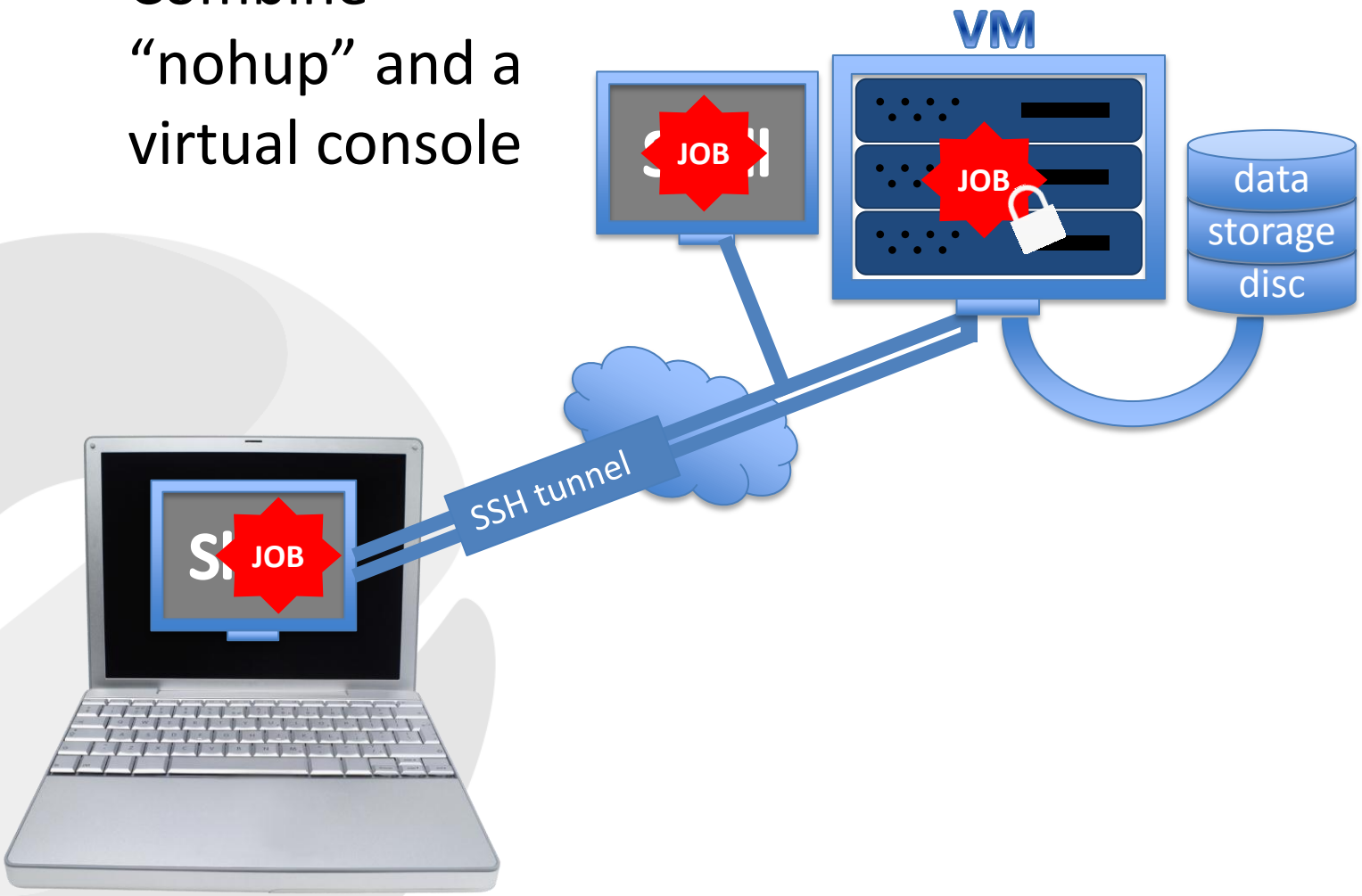
You can connect to the virtual console any time to check on the running job



You can have many virtual consoles, each running different jobs



Combine
“nohup” and a
virtual console



eRSA
Advancing Research Innovation

WORKSHOP: Running a Virtual Machine in the Cloud

GNU PARALLEL

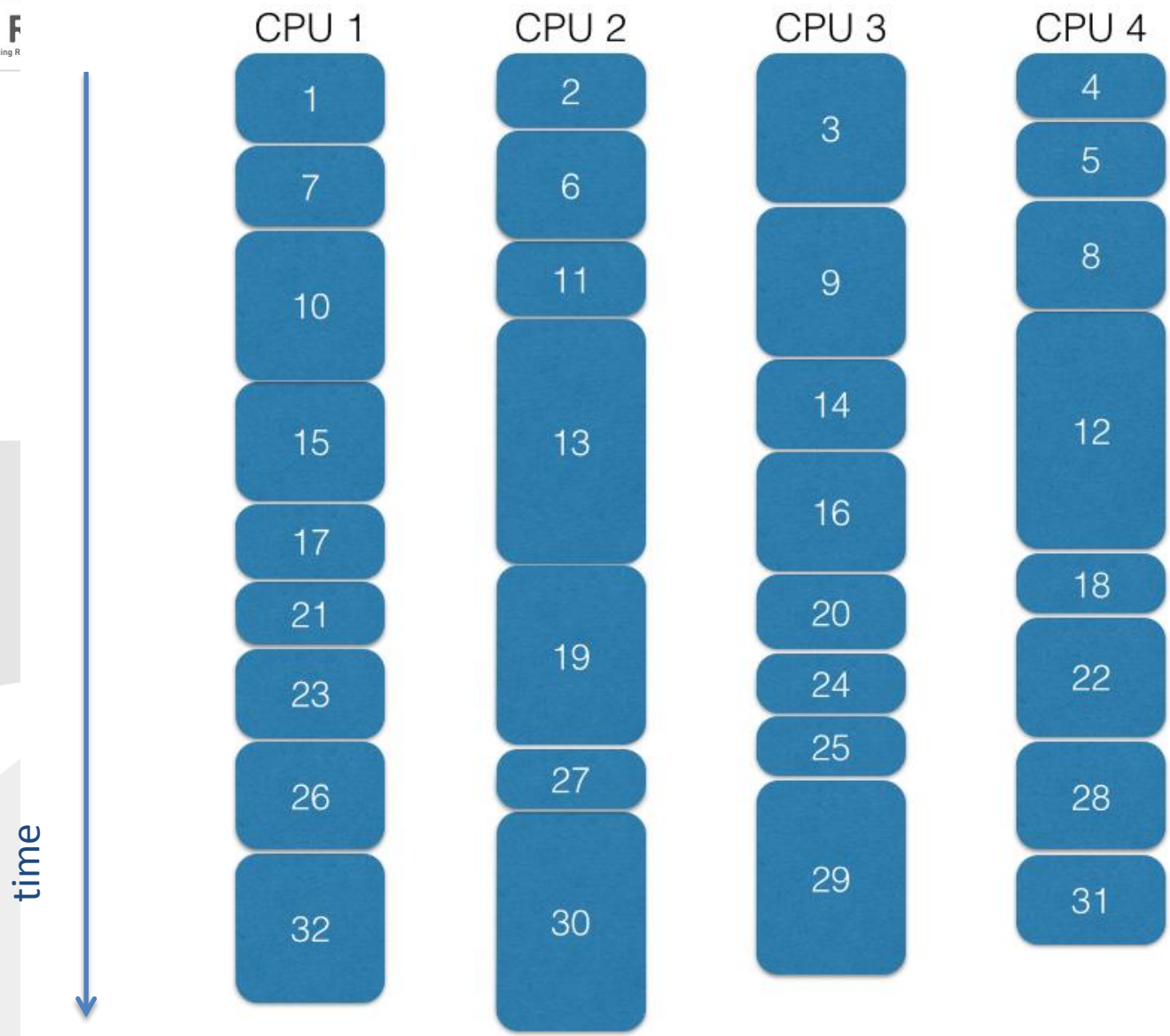
GNU Parallel

- A major advantage of NeCTAR cloud computing, is the power to launch VM's with multiple CPUs
- Multiple CPUs can efficiently process more jobs simultaneously.
- GNU parallel is a command-line utility to manage the distribution of a list of jobs to the available CPU cores.

GNU Parallel

- The GNU parallel utility will allow the user to simultaneously run as many processes as there are CPUs.

- If there are 32 jobs to do and 4 CPUs, parallel will send the first 4 to be done, and as each job finishes a new one will commence.


CPU 1
CPU 2
CPU 3
CPU 4
1
7
10
15
17
21
23
26
32
2
6
11
13
19
27
30
3
9
14
16
20
24
25
29
4
5
8
12
18
22
28
31
time

Upcoming workshops

How to use HPC and Cloud Clusters to enhance your research outcomes
9 March, University of Adelaide

R-Studio in the Cloud
15 April, University of Adelaide

Talk to us after the workshop to register