

The logo for the GPU Technology Conference is located in the top-left corner. It consists of a green rectangular box containing the text "GPU TECHNOLOGY CONFERENCE" in white, sans-serif font. The background of the entire slide is a vibrant, abstract digital grid with glowing lines in shades of blue, green, and purple, creating a sense of depth and connectivity.

**GPU** TECHNOLOGY  
CONFERENCE

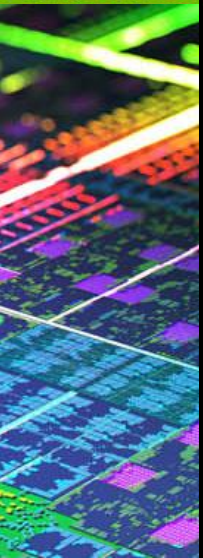
**Building your own GPU Research Cluster  
using Open Source Software stack**

# About the speaker and you

- [Pradeep] is Developer Technology engineer with NVIDIA.
  - I help Customers in parallelizing and optimizing their applications on GPUs.
  - Responsible for GPU evangelism at India and South-East Asia.
- [Audience]
  - Looking for building a research prototype GPU cluster
  - All open-source SW stack for GPU based clusters.

# Outline

- Motivation
- Cluster - Hardware Details
- Cluster Setup - Head Node, Compute Nodes
- Management and Monitoring Snapshots





# Why to build a small GPU based Cluster

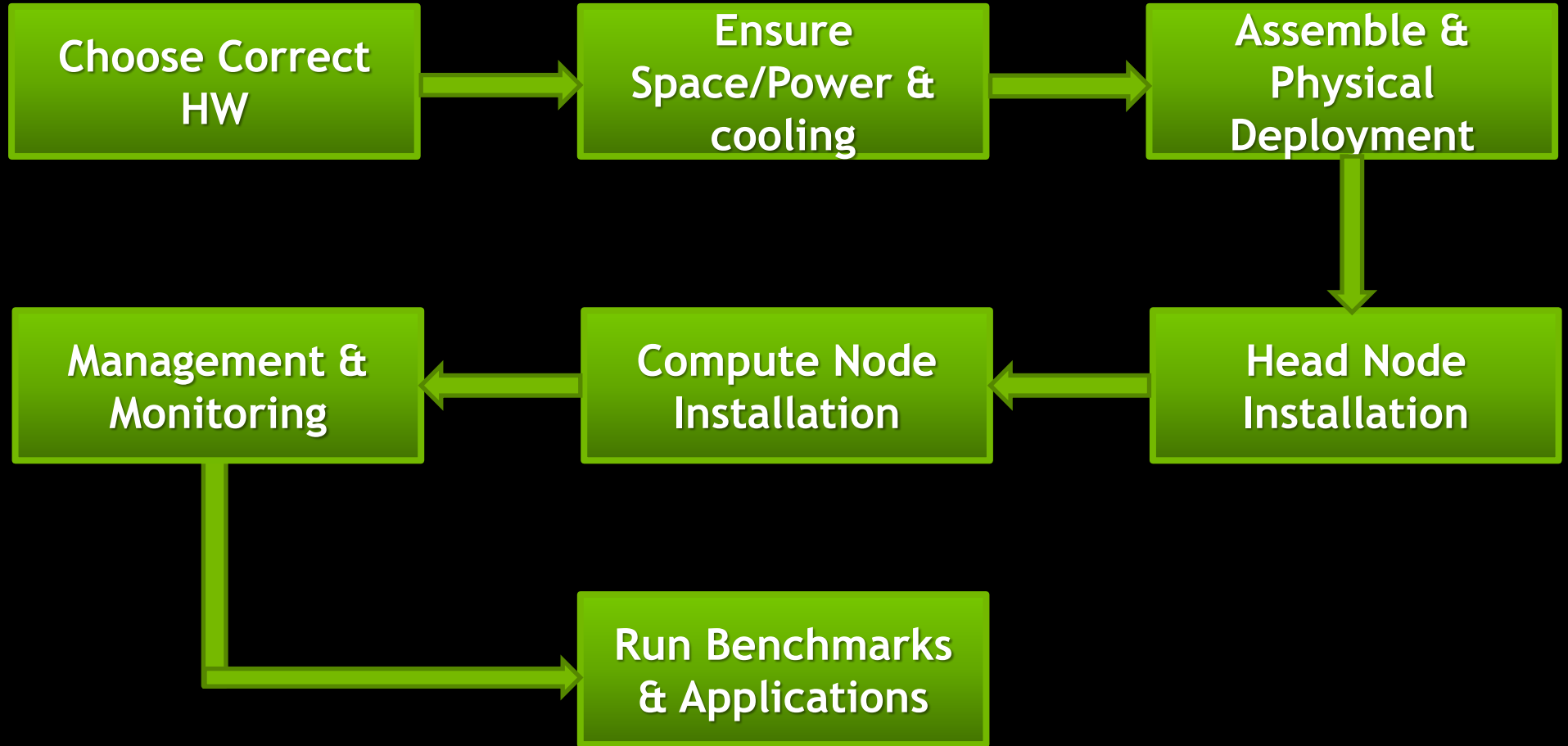
- Get feel of production system and performance estimates
- Port your applications
- GPU and CPU load balancing
- Small investment
- Use it as development platform
- Early experience -> Better readiness

# Today's Focus

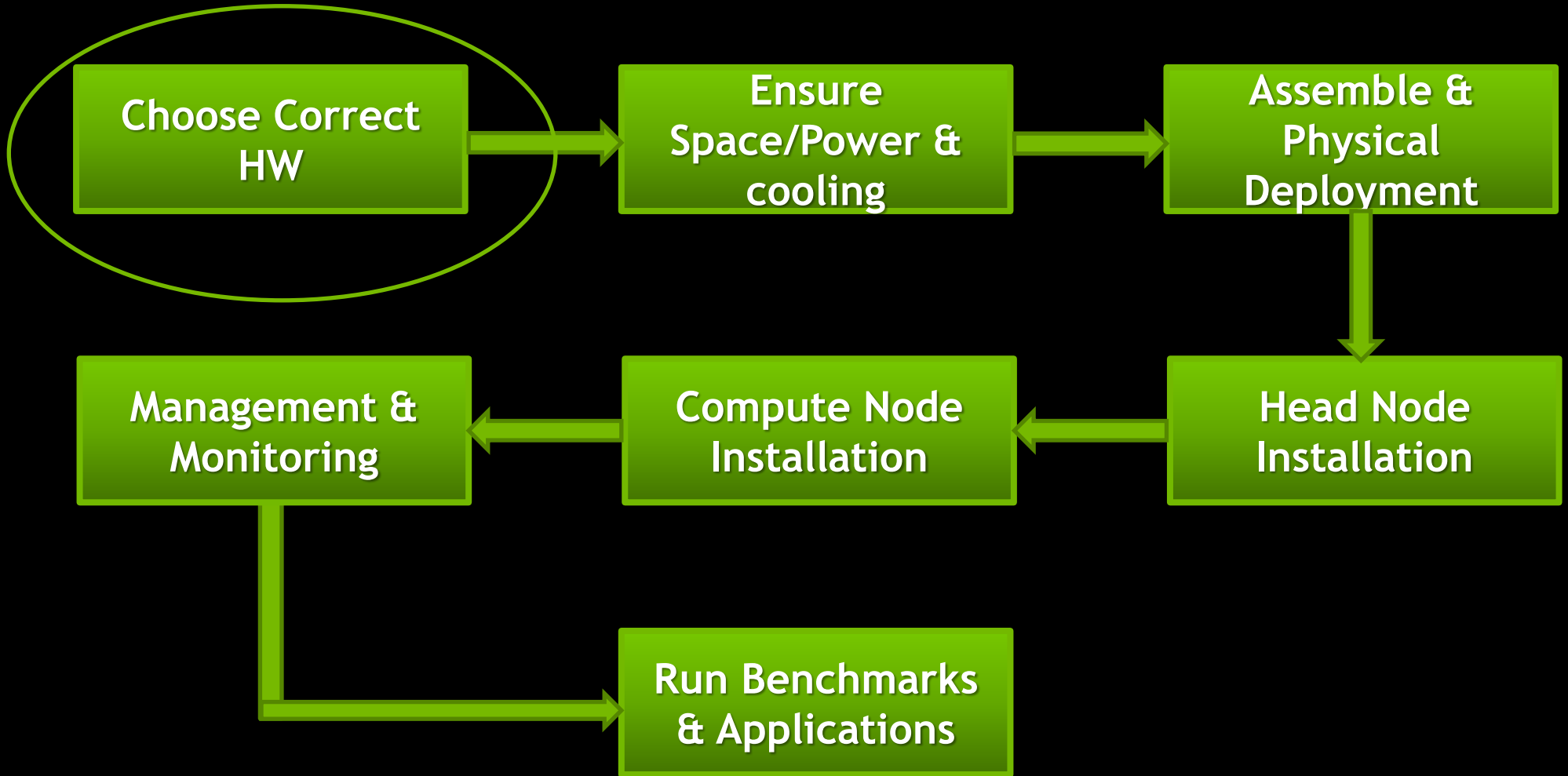
- Trying to build 4 -16 nodes cluster
- Your first GPU based cluster
- You can built in 3 weeks, 2 weeks or less ....
- GPUs - Just add and start using them

Building GPU based Clusters - Very easy, start right now...

# Steps in building GPU based Clusters



# Steps in building GPU based Clusters



## Node HW Details

- CPU Processor
- 2 PCIe x16 wide Gen2/3 connections for Tesla GPUs
- 1 PCIe x8 wide for HCI card for Infiniband
- 2 Network ports
- Min of 16/24 GB DDR3 RAM
- SMPS with required power supply
- 2 x 1 TB HDD

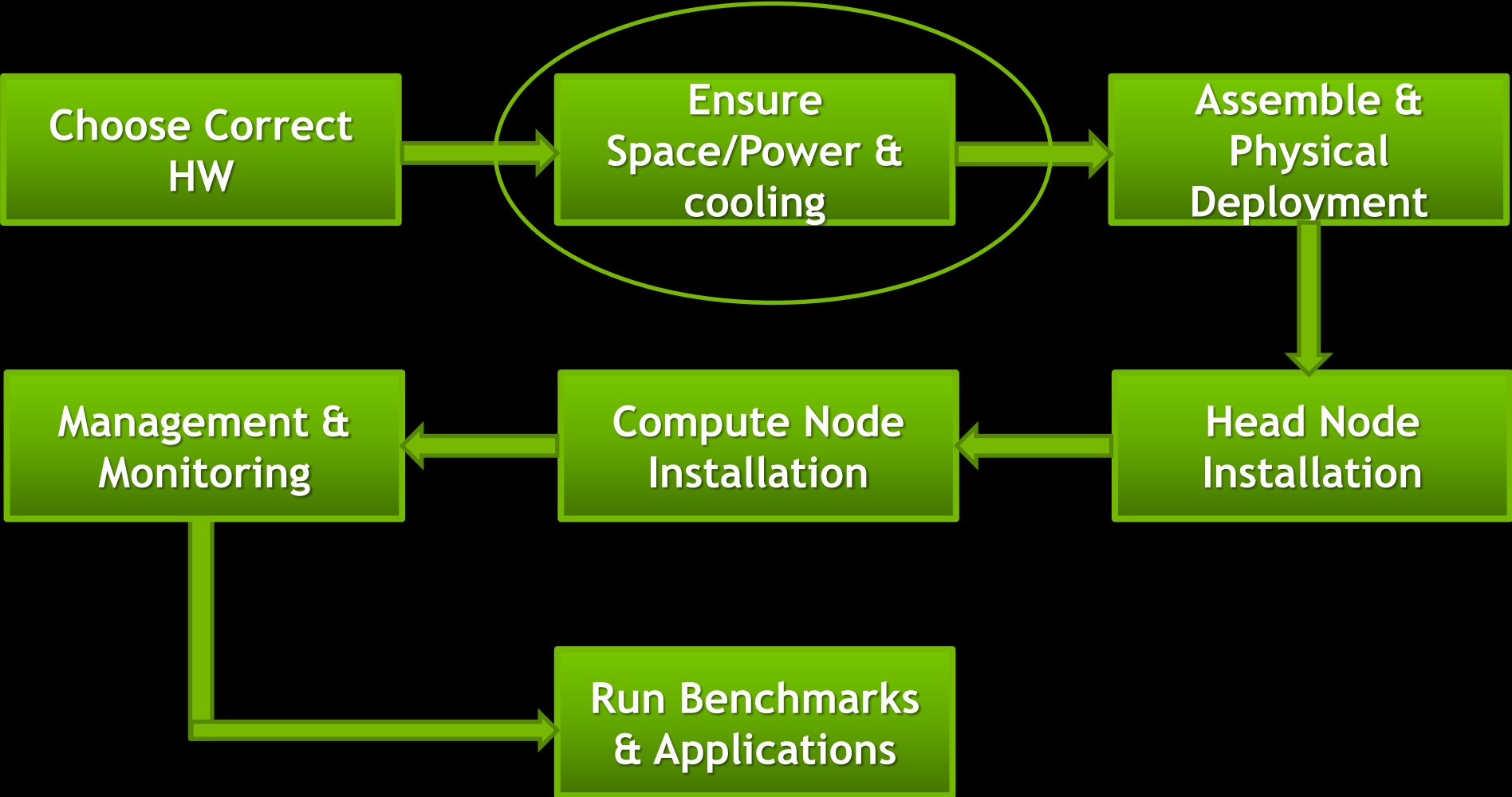


# Tesla™ for Supercomputing

- **Choose the right Form Factor** - Kepler GPUs are available in
  - Workstation products - C Series
  - Server products - M Series
- Different options for adding GPUs
  - Add C series GPUs to existing Workstations
  - Buy a workstation & have C series GPUs
  - Buy servers with M series



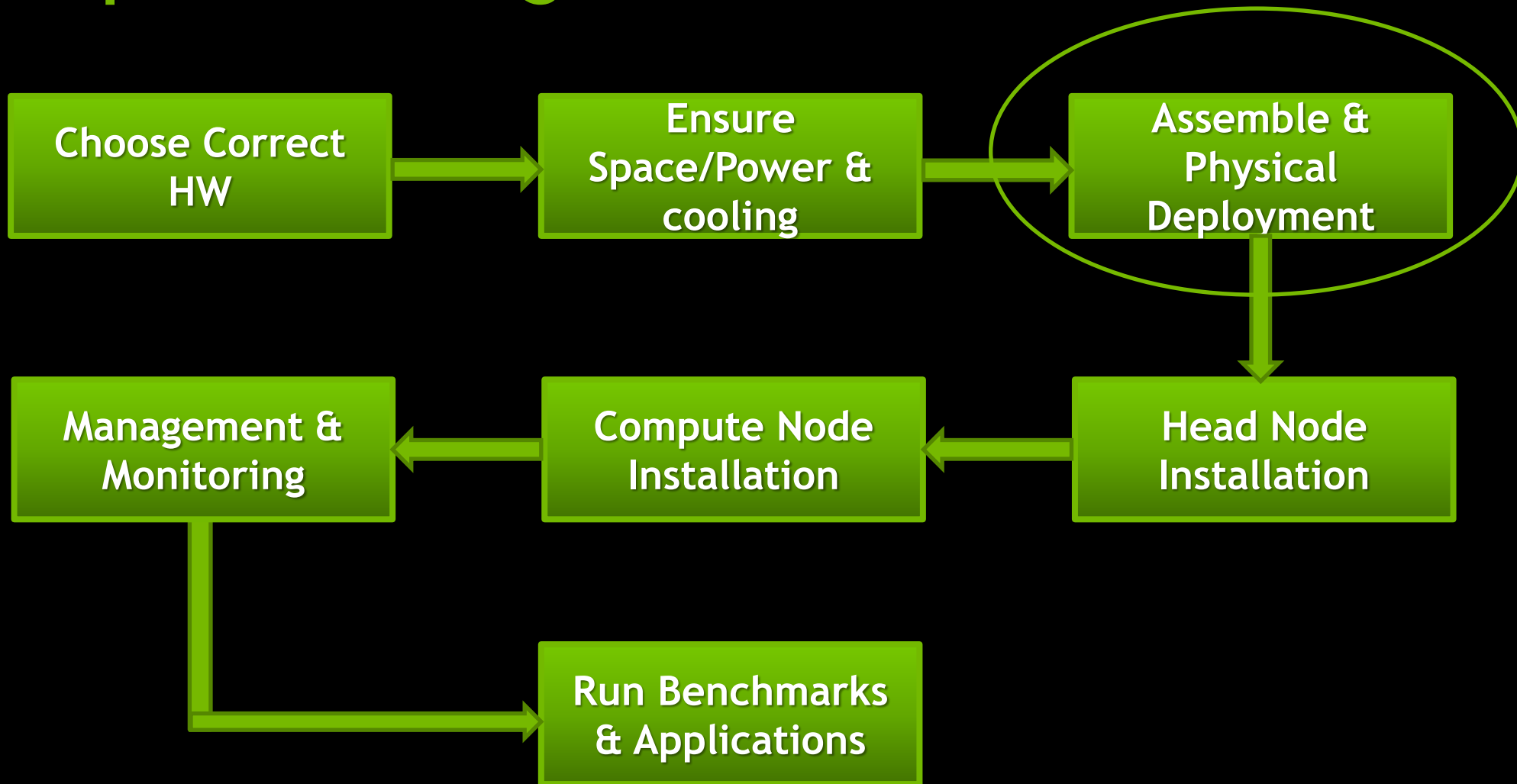
# Steps in building GPU based Clusters



# Other Hardware & real-state

- Space
- Power & Cooling
- Network - Infiniband
- Storage
- Maintenance/repair

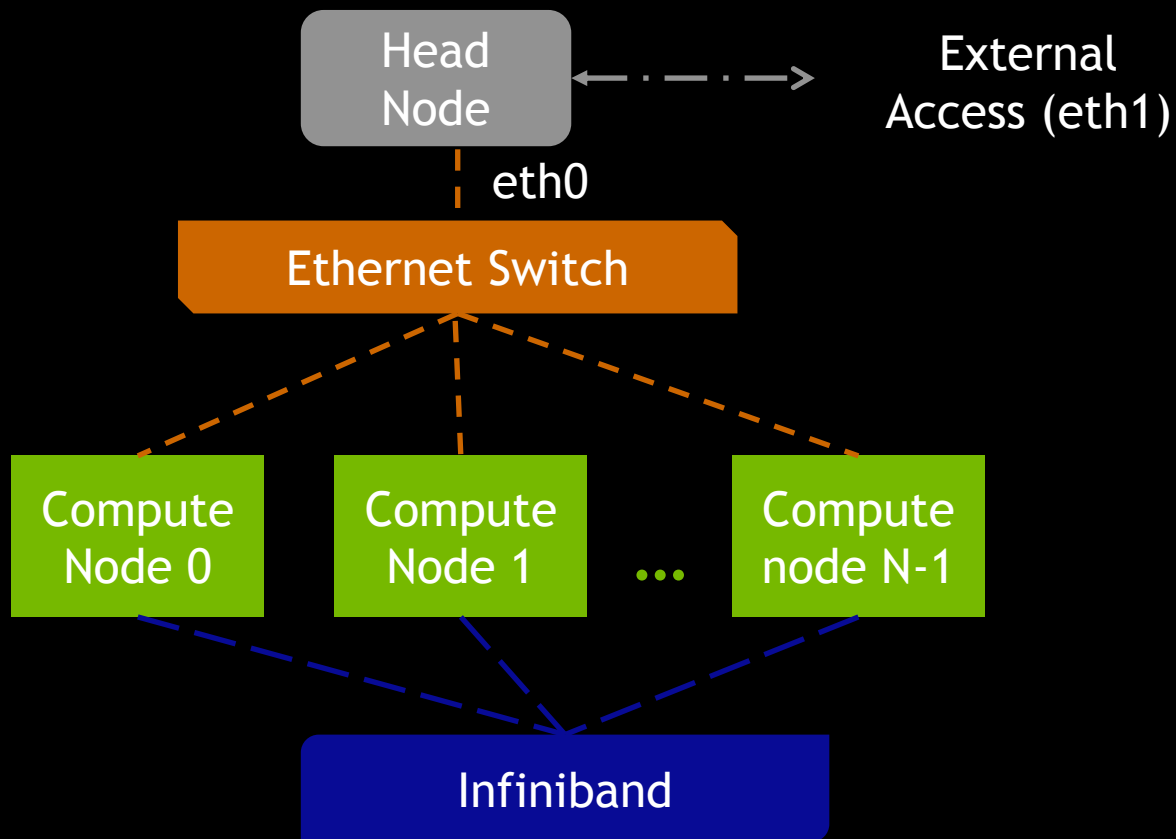
# Steps in building GPU based Clusters



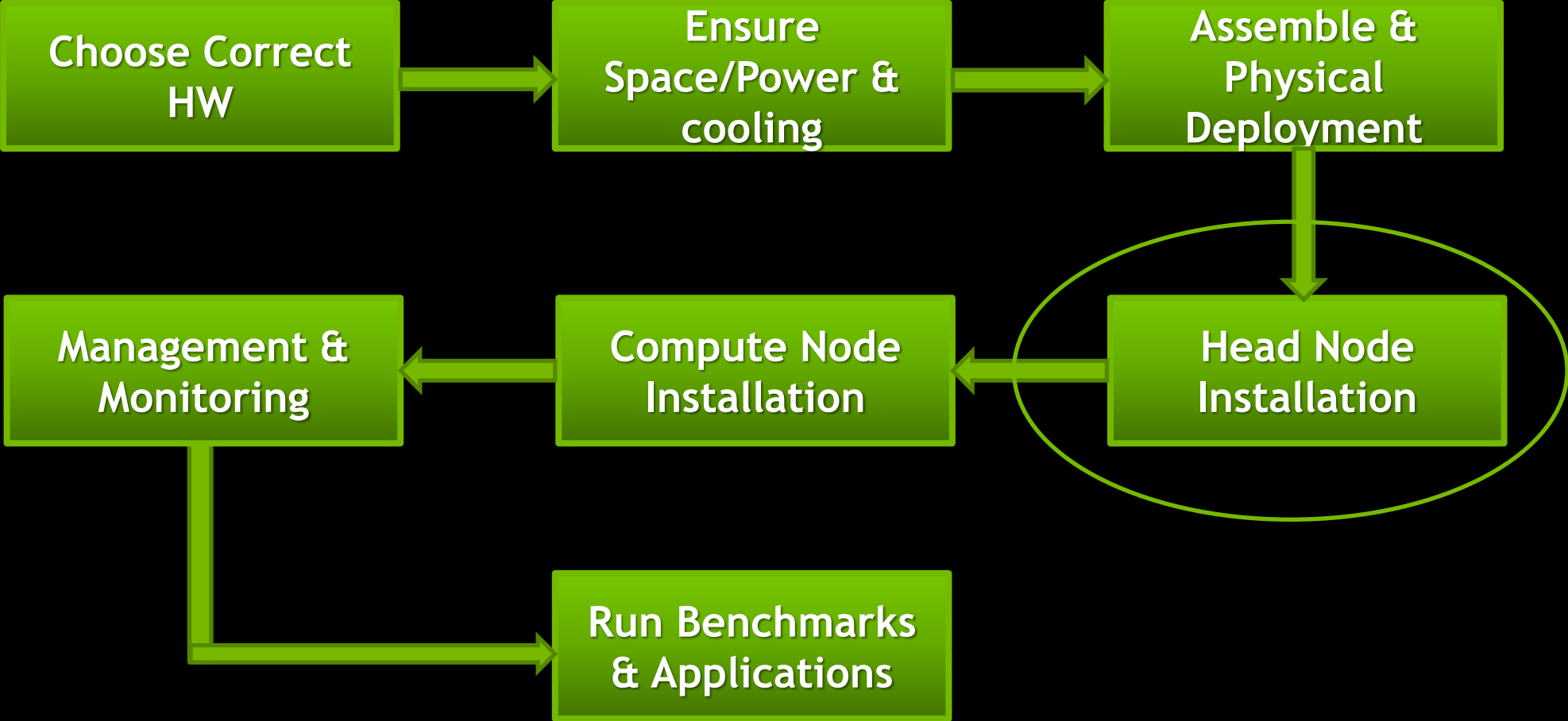


# Setup

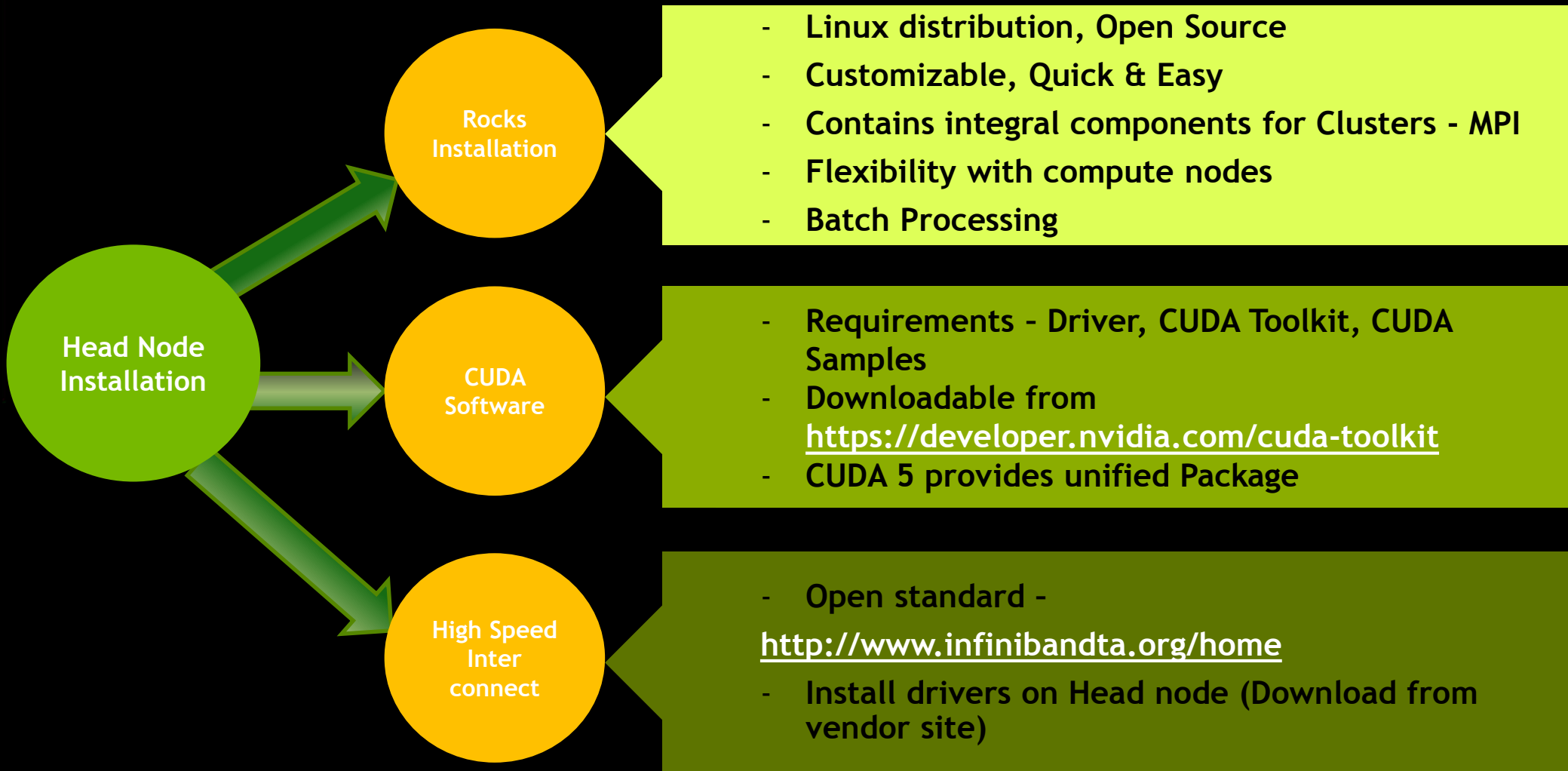
- Physical Deployment of cluster
- Head Node & Compute Nodes connections



# Steps in building GPU based Clusters



# Head Node Installation



# Head Node Installation... contd

Head Node Installation

SW Package Installation

## Nagios Installation -

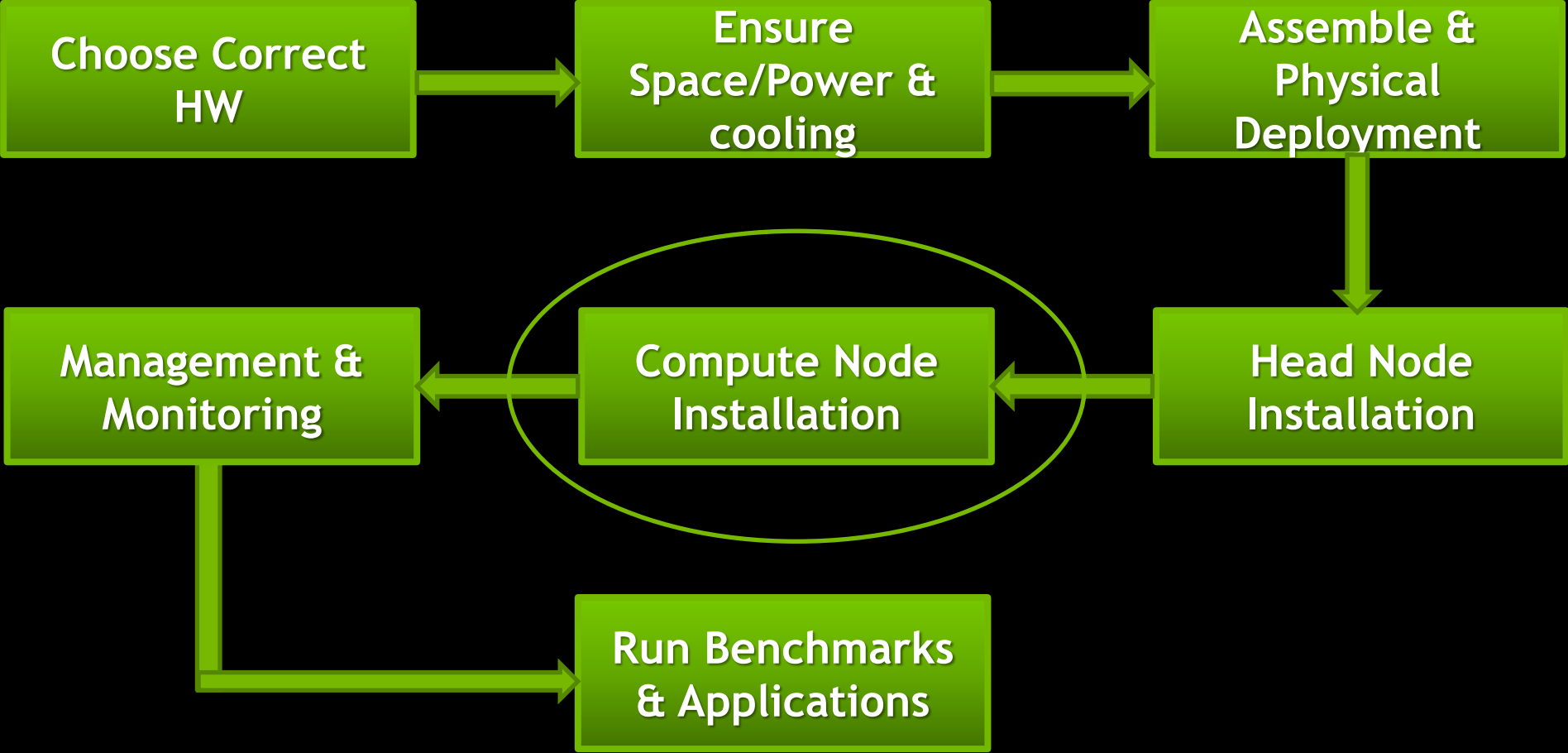
- Monitoring of network services
- Monitoring of host resources
- Web interface and many more ...

## NRPE Installation -

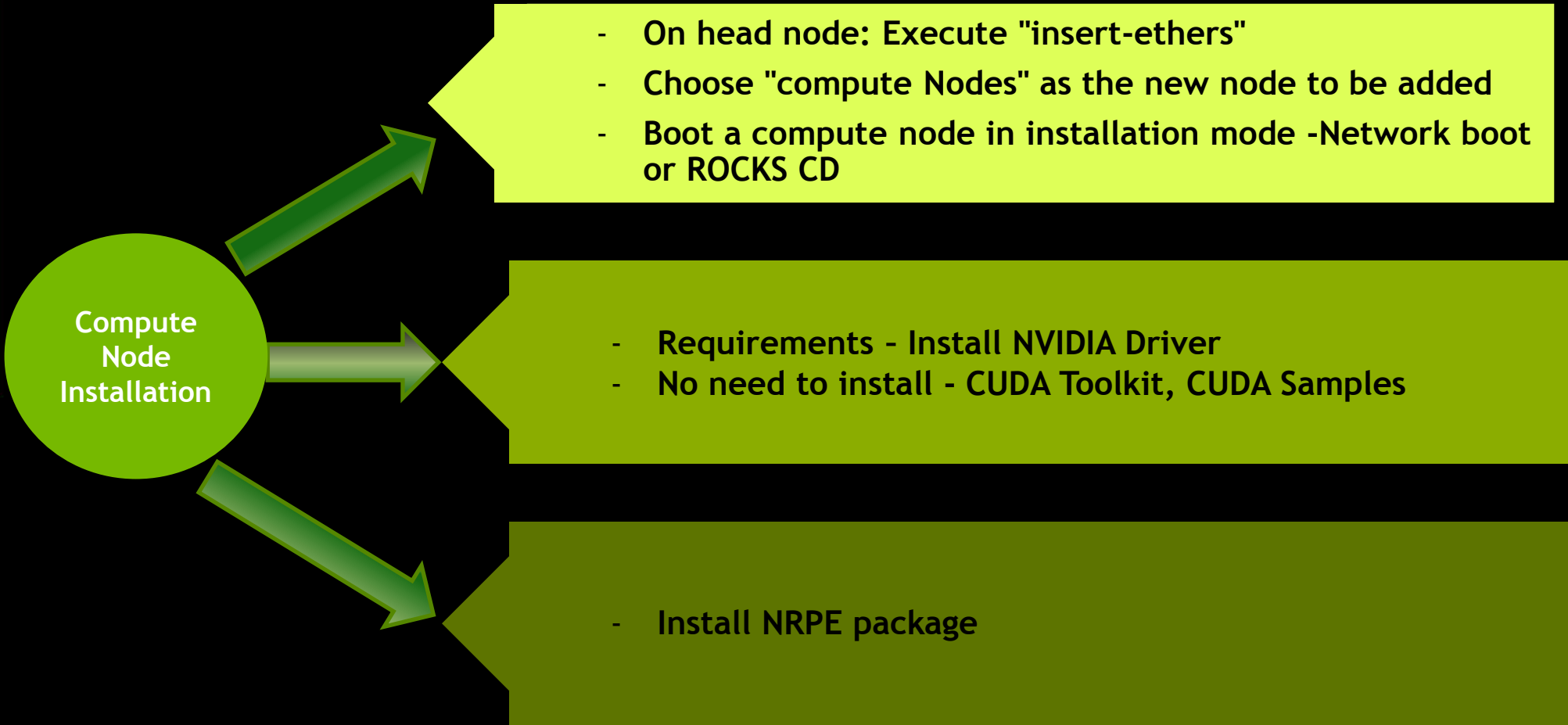
- Execute Nagios plugins on remote machines
- Enables monitoring of local resources



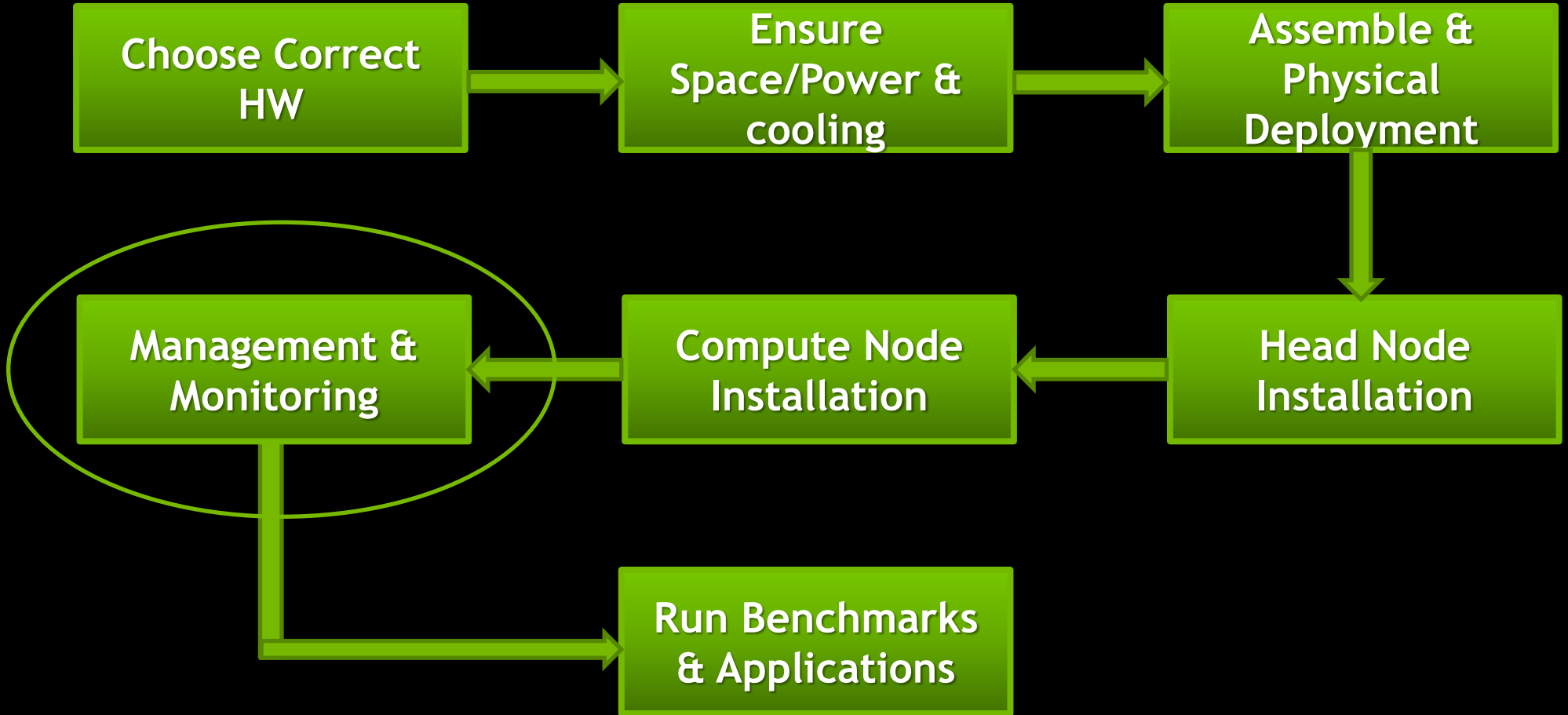
# Steps in building GPU based Clusters



# Compute Node Installation



# Steps in building GPU based Clusters



# System Management for GPU

- **nvidia-smi utility** -
  - Thermal Monitoring Metrics - GPU temperatures, chassis inlet/outlet temperatures
  - System Information- firmware revision, configuration info
  - System State - Fan states, GPU faults, Power system fault etc.
- **nvidia-smi allows you**
  - Different Compute modes - Default/Exclusive/prohibited
  - ECC on/off



# GPU Monitoring

- NVIDIA Provides “TESLA Deployment Kit”
  - Set of tools for better managing Tesla GPUs
  - 2 main components - NVML and nvidia-healthmon
  - <https://developer.nvidia.com/tesla-deployment-kit>
- NVML can be used from Python or Perl
  - NVML - Set of APIs provide state information for GPU monitoring.
- NVML has been integrated into Ganglia gmond.
  - <https://developer.nvidia.com/ganglia-monitoring-system>

# nvidia-healthmon

- Quick health check, Not a full diagnostic
- Suggest remedies to SW and system configuration problems
- Feature Set
  - Basic CUDA and NVML sanity check
  - Diagnosis of GPU failures
  - Check for conflicting drivers
  - Poorly seated GPU detection
  - Check for disconnected power cables
  - ECC error detection and reporting
  - Bandwidth test

# Nagios

**Nagios®**

**Tactical Monitoring Overview**  
 Last Updated: Sat Mar 9 13:50:13 IST 2013  
 Updated every 90 seconds  
 Nagios® Core™ 3.3.1 - www.nagios.org  
 Logged in as nagiosadmin

**Monitoring Performance**

Service Check Execution Time:	0.01 / 4.01 / 1.274 sec
Service Check Latency:	0.01 / 0.25 / 0.131 sec
Host Check Execution Time:	3.01 / 4.01 / 3.757 sec
Host Check Latency:	0.07 / 0.09 / 0.085 sec
# Active Host / Service Checks:	4 / 29
# Passive Host / Service Checks:	0 / 0

**Network Health**

Host Health:

Service Health:

**Monitoring of Nodes** (Annotation pointing to Hosts section)

**Monitoring of Services** (Annotation pointing to Services section)

**General**

- Home
- Documentation

**Current Status**

- Tactical Overview
- Map
- Hosts
- Services
- Host Groups
  - Summary
  - Grid
- Service Groups
  - Summary
  - Grid
- Problems
  - Services (Unhandled)
  - Hosts (Unhandled)
  - Network Outages

Quick Search:

**Reports**

- Availability
- Trends
- Alerts
  - History
  - Summary
  - Histogram
- Notifications
- Event Log

**System**

- Comments
- Downtime

**Network Outages**

0 Outages

**Hosts**

1 Down    0 Unreachable    3 Up    0 Pending

1 Unhandled Problems

**Services**

6 Critical    0 Warning    0 Unknown    23 Ok    0 Pending

6 on Problem Hosts

**Monitoring Features**

Flap Detection	Notifications	Event Handlers	Active Checks	Passive Checks
✓ All Services Enabled	✓ 5 Services Disabled	✓ All Services Enabled	✓ All Services Enabled	✓ All Services Enabled
No Services Flapping	All Hosts Enabled	All Hosts Enabled	All Hosts Enabled	All Hosts Enabled
All Hosts Enabled				
No Hosts Flapping				

# Nagios - GPU Memory Usage

**Nagios**

**Service Information**  
 Last Updated: Sat Mar 9 14:26:13 IST 2013  
 Updated every 90 seconds  
 Nagios® Core™ 3.3.1 - www.nagios.org  
 Logged in as nagiosadmin

**Service**  
**GPU Memory**  
 On Host localhost (localhost)  
 Member of No servicegroups.  
 127.0.0.1

## Monitoring of GPU Memory

**Service State Information**

Current Status: **OK** (for 10d 2h 16m 47s)  
 Status Information: OK: Memory Free : 100 %  
 Performance Data: 1/4 (HARD state)  
 Current Attempt: 03-09-2013 14:24:05  
 Last Check Time: ACTIVE  
 Check Type: 0.010 / 0.225 seconds  
 Check Latency / Duration: 03-09-2013 14:29:05  
 Next Scheduled Check: 02-19-2013 12:09:26  
 Last State Change: N/A (notification 0)  
 Last Notification: NO (0.00% state change)  
 Is This Service Flapping? NO  
 In Scheduled Downtime? NO  
 Last Update: 03-09-2013 14:26:05 ( 0d 0h 0m 8s ago)

Active Checks: **ENABLED**  
 Passive Checks: **ENABLED**  
 Obsessing: **ENABLED**  
 Notifications: **ENABLED**  
 Event Handler: **ENABLED**  
 Flap Detection: **ENABLED**


**General**  
 Home  
 Documentation

**Current Status**  
 Tactical Overview  
 Map  
 Hosts  
 Services  
 Host Groups  
 • Summary  
 • Grid  
 Service Groups  
 • Summary  
 • Grid  
 Problems  
 • Services (Unhandled)  
 • Hosts (Unhandled)  
 • Network Outages

Quick Search:

**Reports**  
 Availability  
 Trends  
 Alerts  
 • History  
 • Summary  
 • Histogram  
 Notifications  
 Event Log

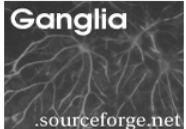
# Research Cluster



## PDC-Rocks-Cluster Report for Sat, 09 Mar 2013 14:24:12 +0530

Last  or from  to

Metric  Sorted



[Physical View](#)

[Grid > PDC-Rocks-Cluster](#) >

---

### Overview of PDC-Rocks-Cluster

CPU's Total: **20**

Hosts up: **3**

Hosts down: **0**

Current Load Avg (15, 5, 1m):  
**0%, 1%, 1%**


Avg Utilization (last month):  
**1%**

Localtime:  
**2013-03-09 14:24**

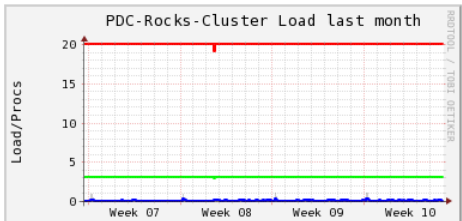
---

Cluster Load Percentages

0-25 (100.00%)

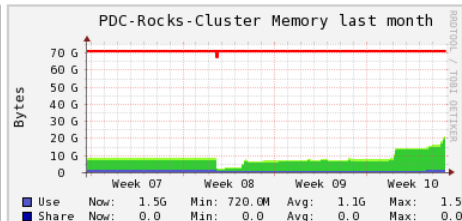


#### PDC-Rocks-Cluster Load last month



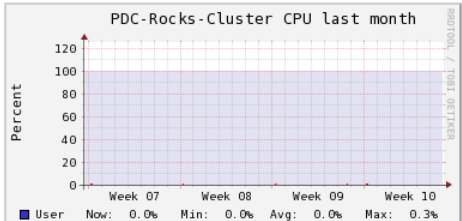
1-min	Now: 118.1m	Min: 22.2m	Avg: 128.9m	Max: 983.0m
Nodes	Now: 3.0	Min: 2.8	Avg: 3.0	Max: 3.0
CPU's	Now: 20.0	Min: 19.0	Avg: 20.0	Max: 20.0
Procs	Now: 59.8m	Min: 0.0	Avg: 49.4m	Max: 365.2m

#### PDC-Rocks-Cluster Memory last month



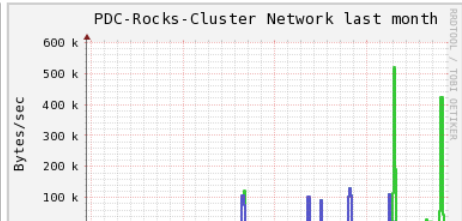
Use	Now: 1.5G	Min: 720.0M	Avg: 1.1G	Max: 1.5G
Share	Now: 0.0	Min: 0.0	Avg: 0.0	Max: 0.0
Cache	Now: 18.3G	Min: 1.1G	Avg: 6.4G	Max: 18.3G
Buffer	Now: 968.1M	Min: 97.9M	Avg: 795.2M	Max: 968.1M
Swap	Now: 0.0	Min: 0.0	Avg: 0.0	Max: 0.0
Total	Now: 70.6G	Min: 67.0G	Avg: 70.6G	Max: 70.6G

#### PDC-Rocks-Cluster CPU last month



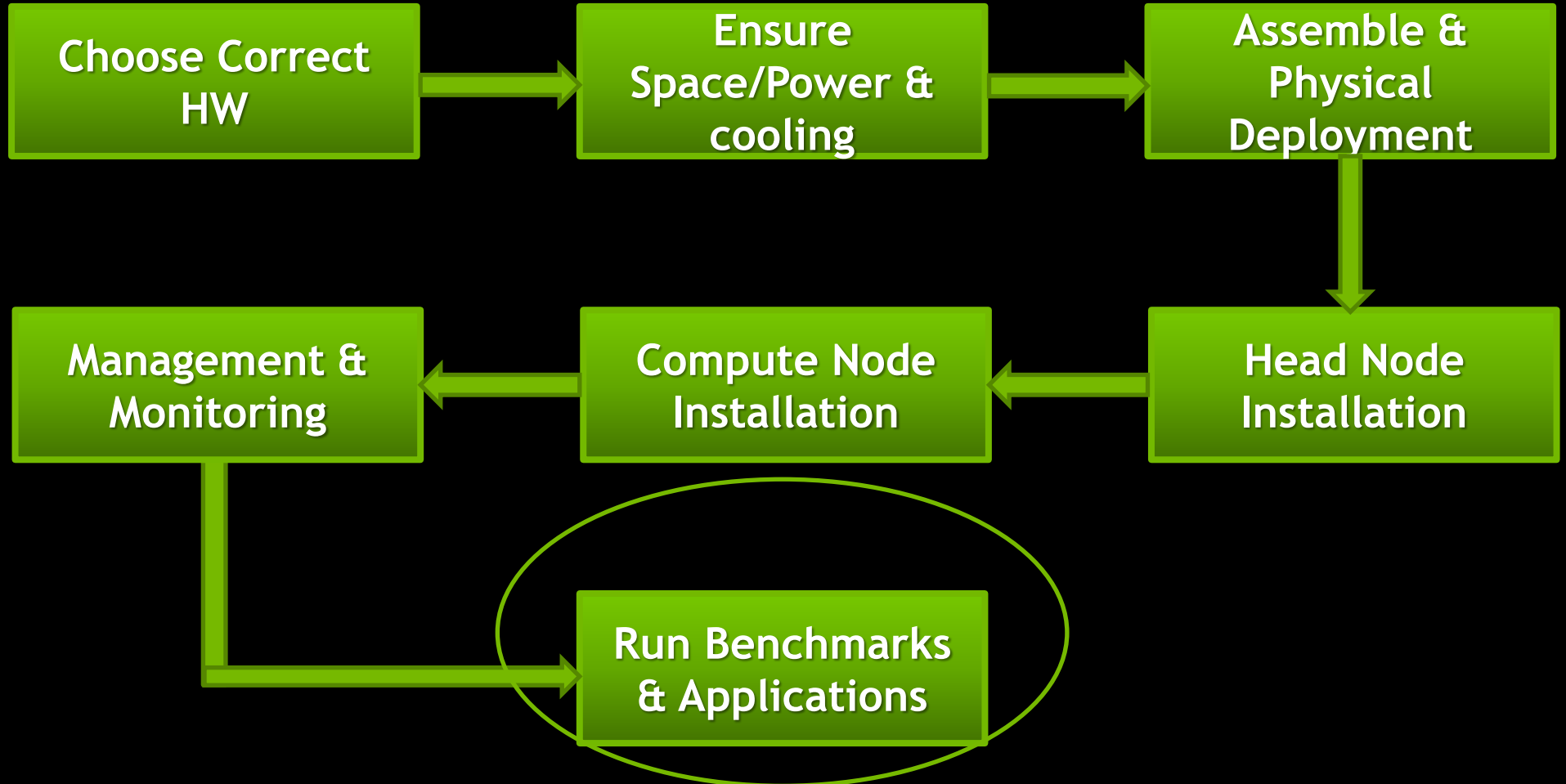
User	Now: 0.0%	Min: 0.0%	Avg: 0.0%	Max: 0.3%
Nice	Now: 0.0%	Min: 0.0%	Avg: 0.0%	Max: 0.0%
System	Now: 0.3%	Min: 0.2%	Avg: 0.3%	Max: 1.1%
Wait	Now: 0.0%	Min: 0.0%	Avg: 0.0%	Max: 0.1%
Idle	Now: 99.7%	Min: 98.7%	Avg: 99.7%	Max: 99.8%

#### PDC-Rocks-Cluster Network last month



In	Now: 476.2	Min: 472.3	Avg: 10.1k	Max: 517.3k
Out	Now: 622.8	Min: 616.1	Avg: 4.4k	Max: 127.4k

# Steps in building GPU based Clusters





# Benchmarks

- GPUs

- devicequery
- Bandwidth Test

- Infiniband

- Bandwidth and latency test
- `<MPI Install PATH>/tests/osu_benchmarks-3.1.1`
- Use Open Source CUDA-aware MPI implementation like `MVAPICH2`

- Application

- LINPACK

Questions ?  
See you at GTC 2014

