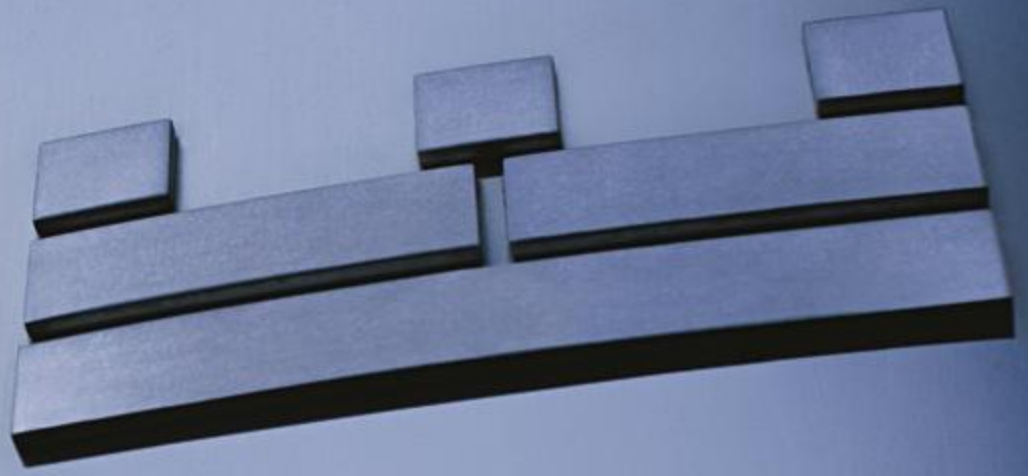
CITADEL



GPGPU Option Models

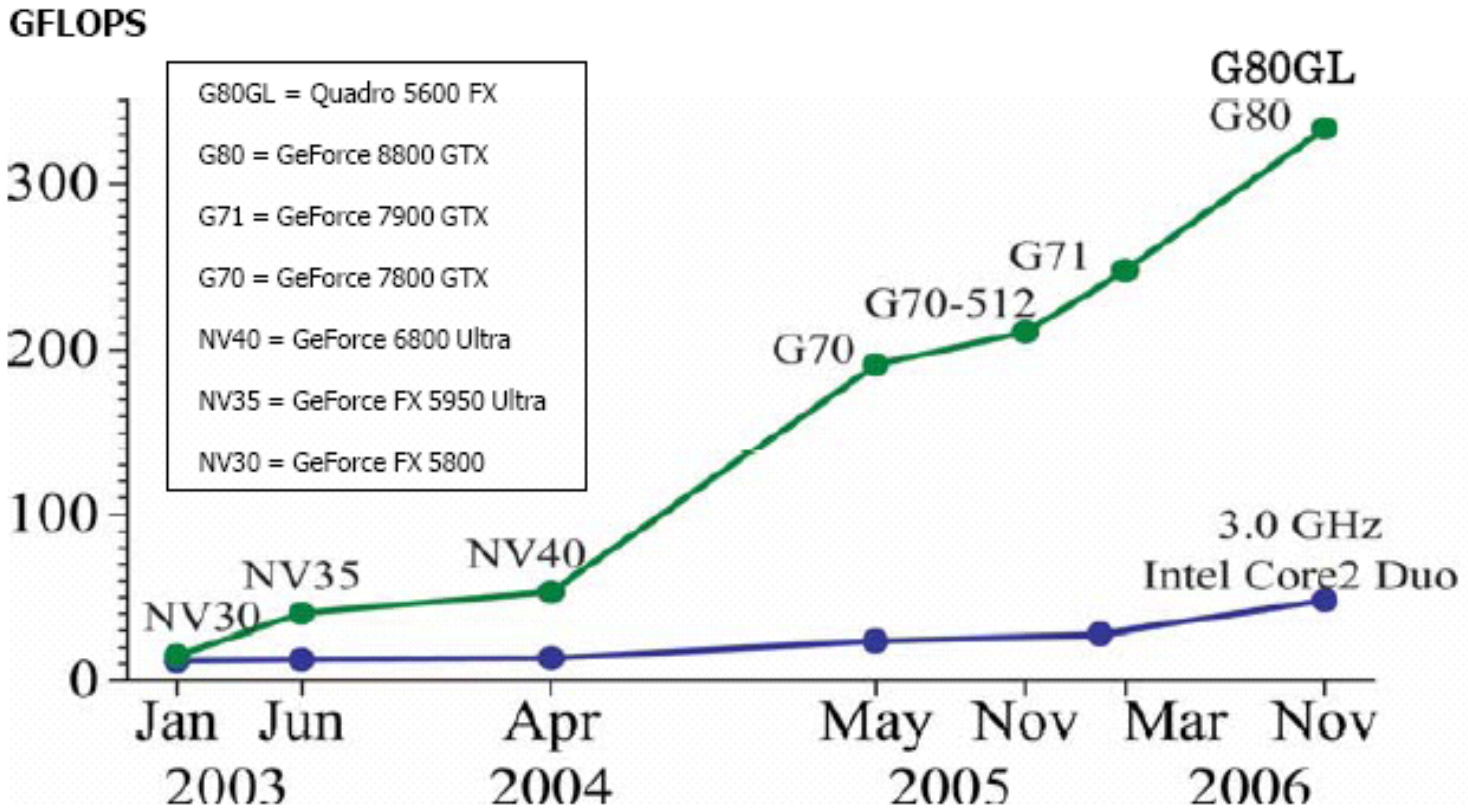
Integrated Into Existing Infrastructure via Virtualization

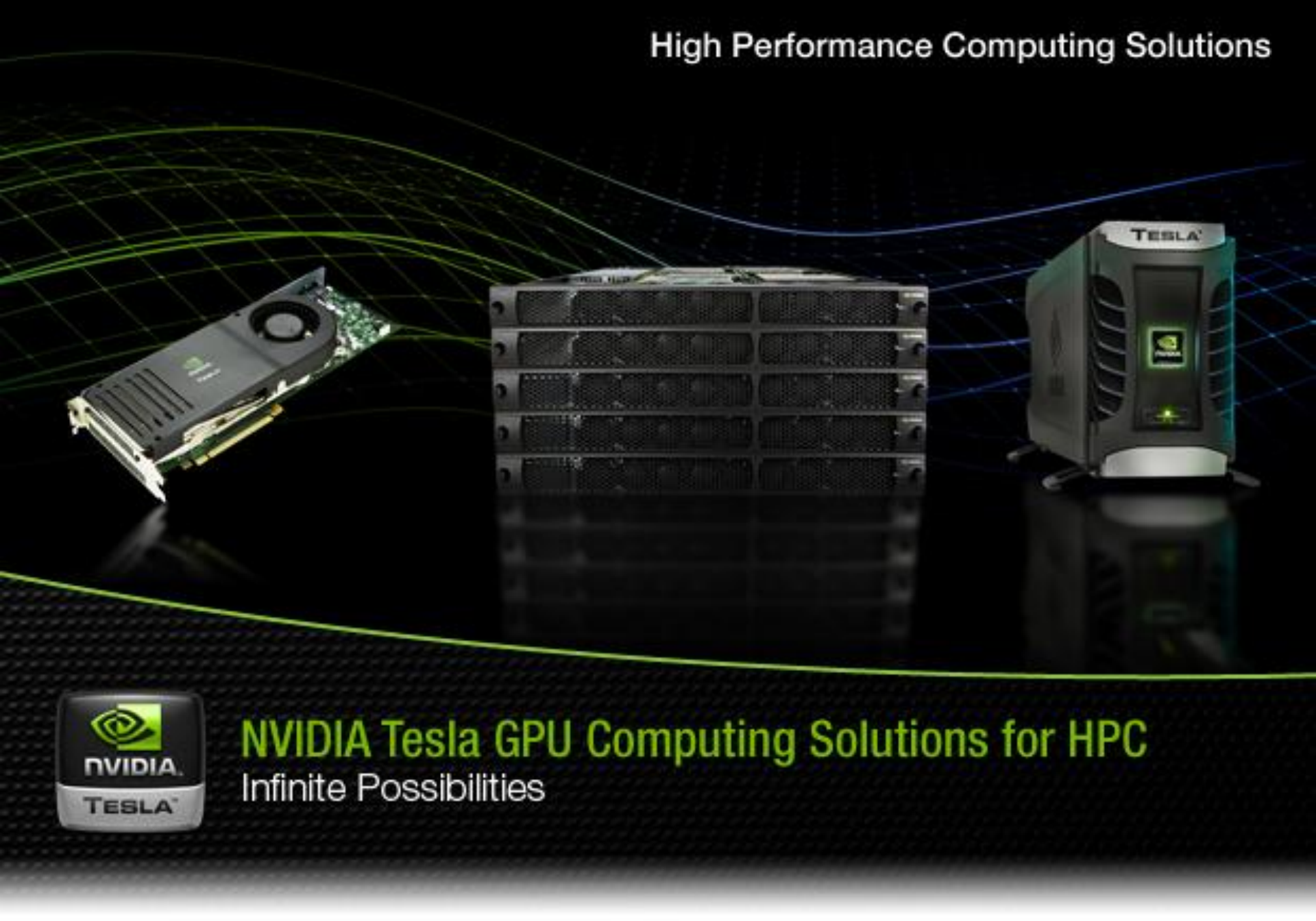
September 17, 2010

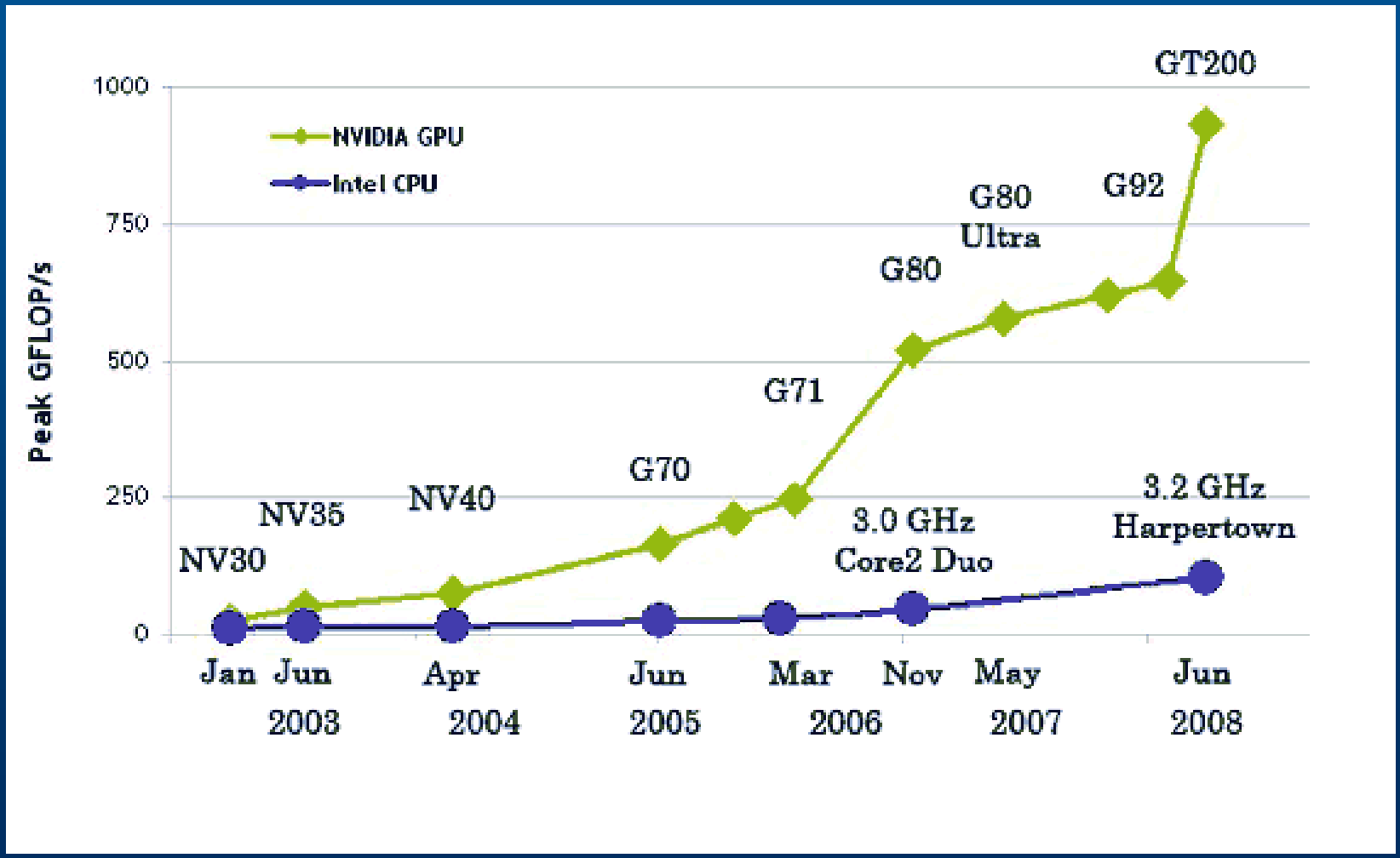
Overview

- Citadel's GPGPU Pilot Project
- Citadel's "Valuation" infrastructure
- Virtualized GPUs
- Demo

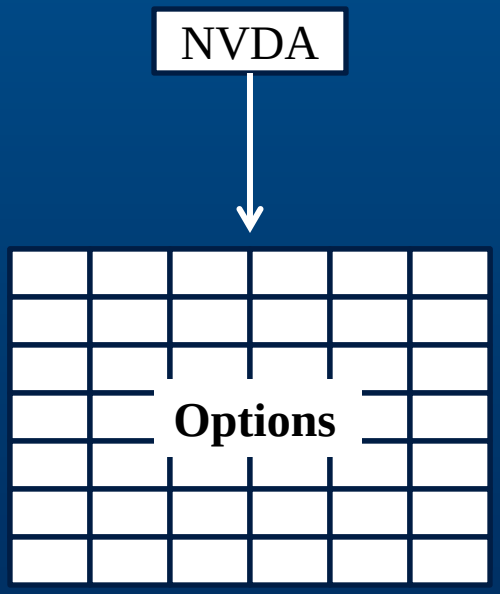
GPGPU Computing


GFLOPS
G80GL = Quadro 5600 FX
G80 = GeForce 8800 GTX
G71 = GeForce 7900 GTX
G70 = GeForce 7800 GTX
NV40 = GeForce 6800 Ultra
NV35 = GeForce FX 5950 Ultra
NV30 = GeForce FX 5800
300
200
100
0
G80GL
G80
G71
G70-512
G70
NV40
NV35
NV30
3.0 GHz
Intel Core2 Duo
Jan
Jun
Apr
May
Nov
Mar
Nov
2003
2004
2005
2006

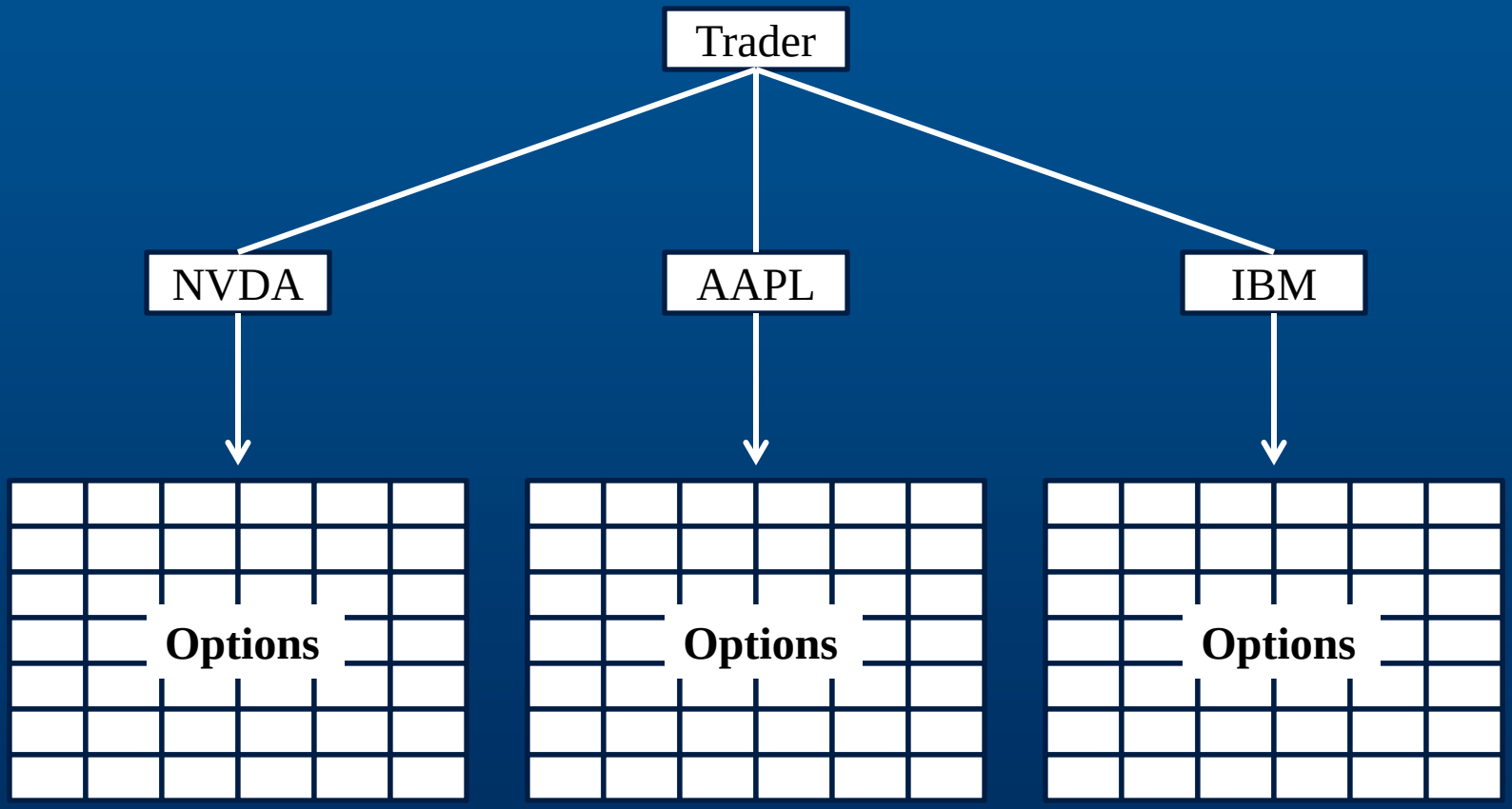

High Performance Computing Solutions
TESLA
NVIDIA
TESLA
NVIDIA Tesla GPU Computing Solutions for HPC
Infinite Possibilities


Peak GFLOP/s
1000
750
500
250
0
NVIDIA GPU
Intel CPU
GT200
G92
G80
Ultra
G80
G71
G70
NV40
NV35
NV30
3.2 GHz
Harpertown
3.0 GHz
Core2 Duo
Jan
Jun
Apr
Jun
Mar
Nov
May
Jun
2003
2004
2005
2006
2007
2008

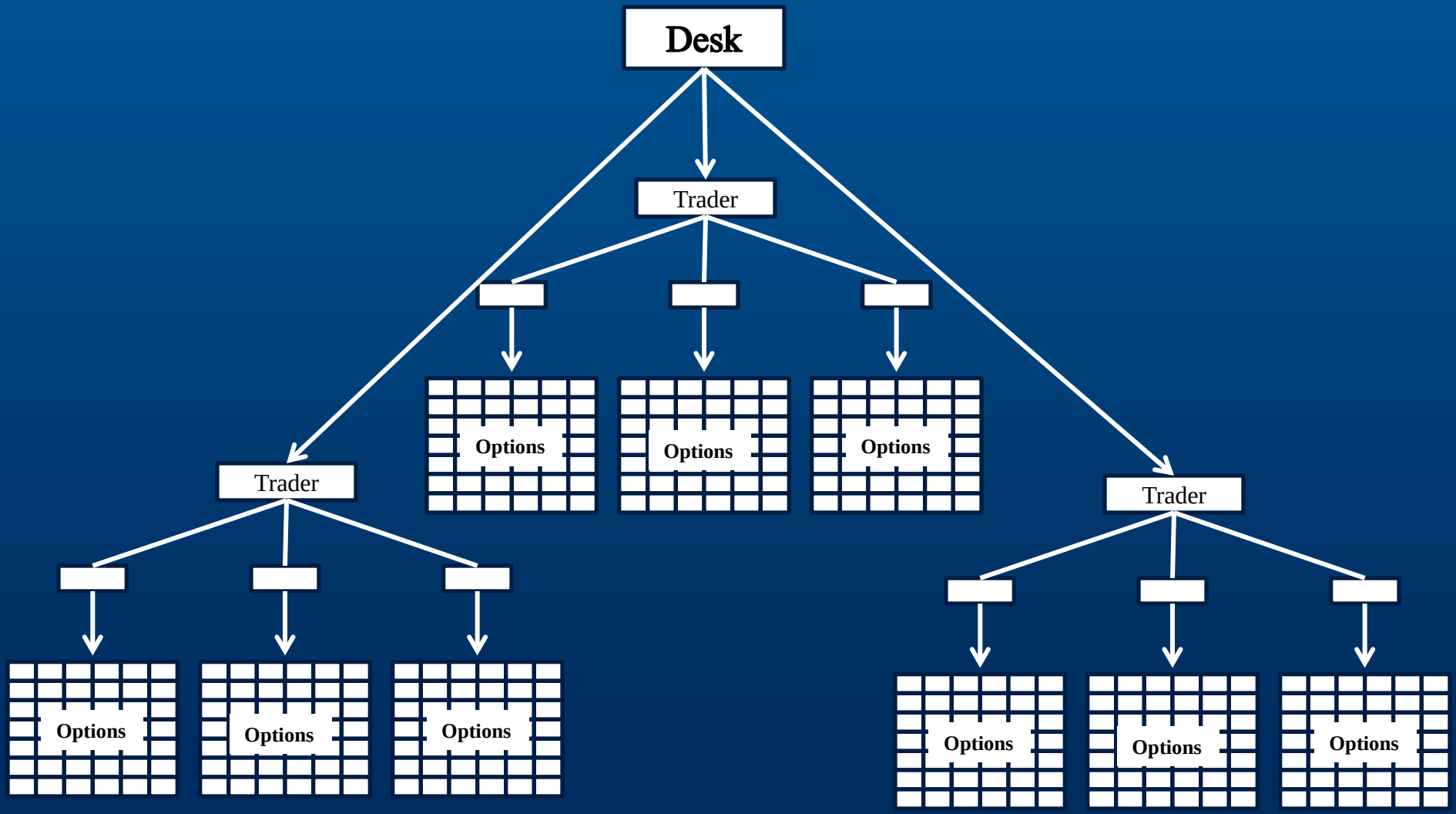
Number of Theoretical Value Calculations



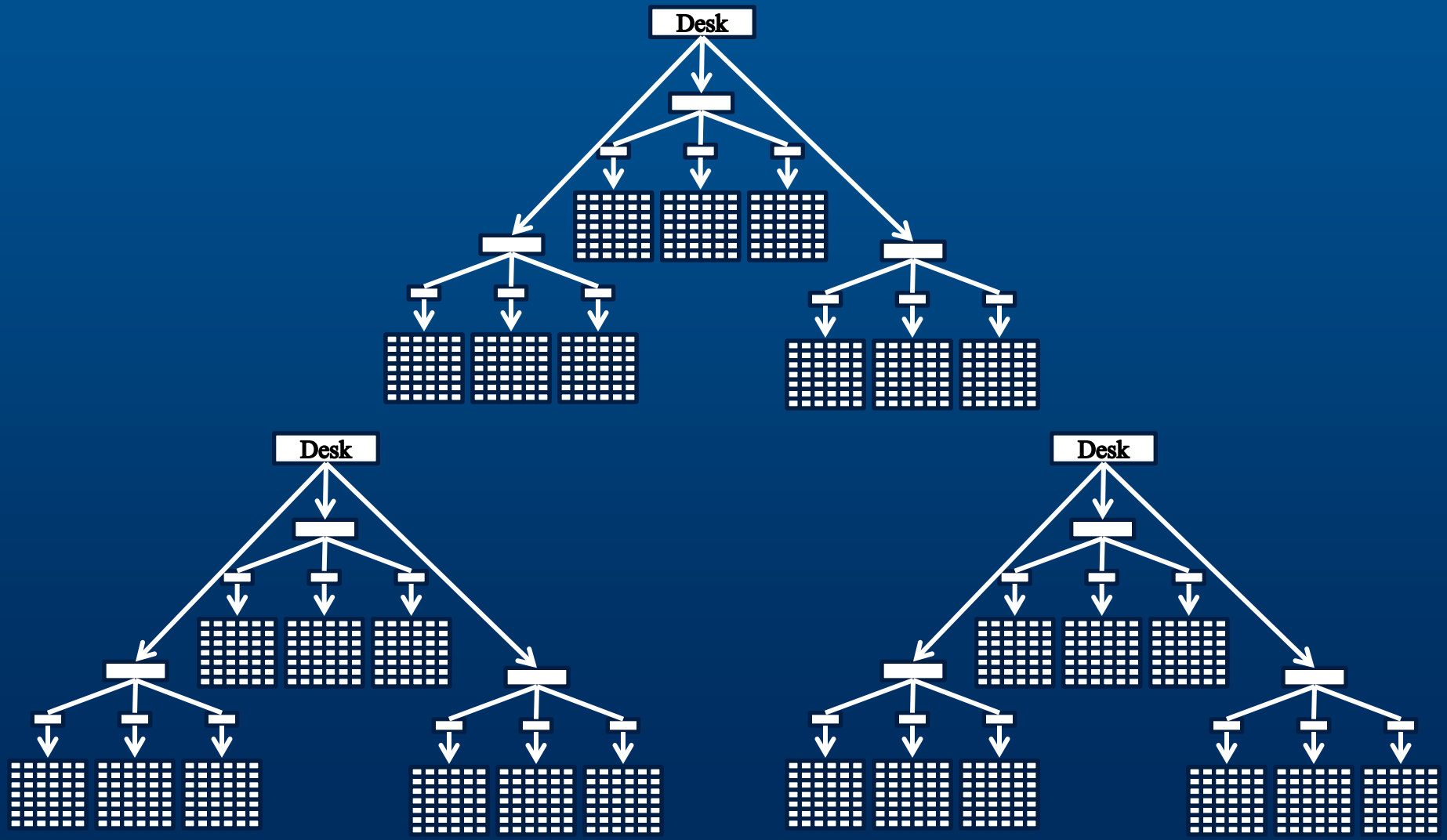
Number of Theoretical Value Calculations



Number of Theoretical Value Calculations

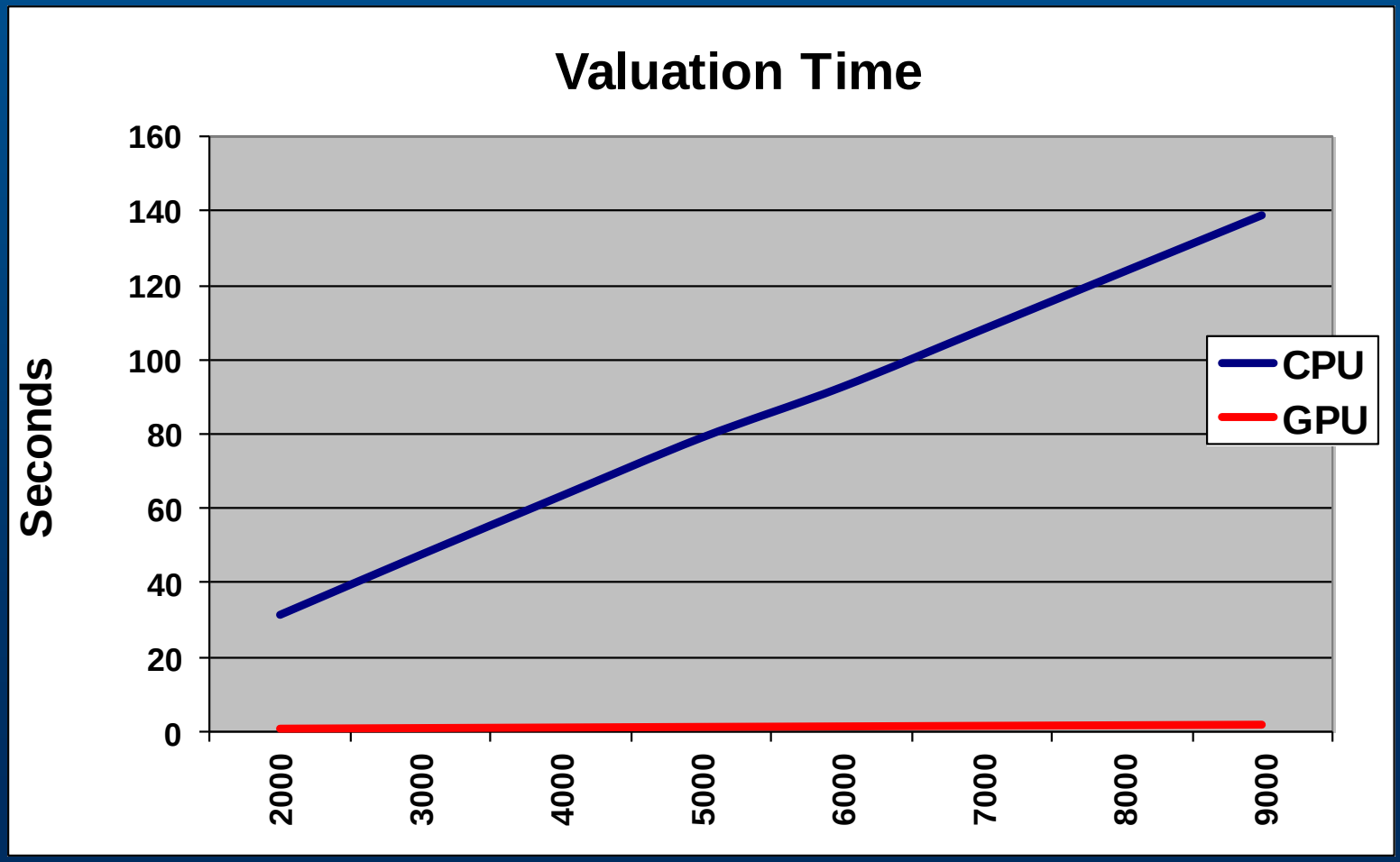


Number of Theoretical Value Calculations



Global Equities Option Model (Theoretical value)

Trinomial Option Model



Global Equities Option Model (Theoretical value)

Trinomial Tree Option Model

- We have been up and running in prod for the past year

Global Equities Option Model (Theoretical value)

Trinomial Tree Option Model

- We have been up and running in prod for the past year
- We have performed over 3 billon calculations

Global Equities Option Model (Theoretical value)

Trinomial Tree Option Model

- We have been up and running in prod for the past year
- We have performed over 3 billon calculations
- We are currently utilizing 14 GPUs

Global Equities Option Model (Theoretical value)

Trinomial Tree Option Model

- We have been up and running in prod for the past year
- We have performed over 3 billon calculations
- We are currently utilizing 14 GPUs
- We are saving $425,000 in hardware cost per year

CPU Hardware Cost

A transparent breakdown

- One time costs
  - The server hardware costs $728 per core
  - The rack and network gear costs $4 per core
  - The build out cost would be $5 per core
- Recurring costs
  - SA costs are around $111 per core
  - Datacenter space costs around $94 per core

CPU Hardware Cost

A transparent breakdown

- $737 per core (One time cost)
- $617 per core (3 years of recurring costs)

- $1,350 per core every 3 years

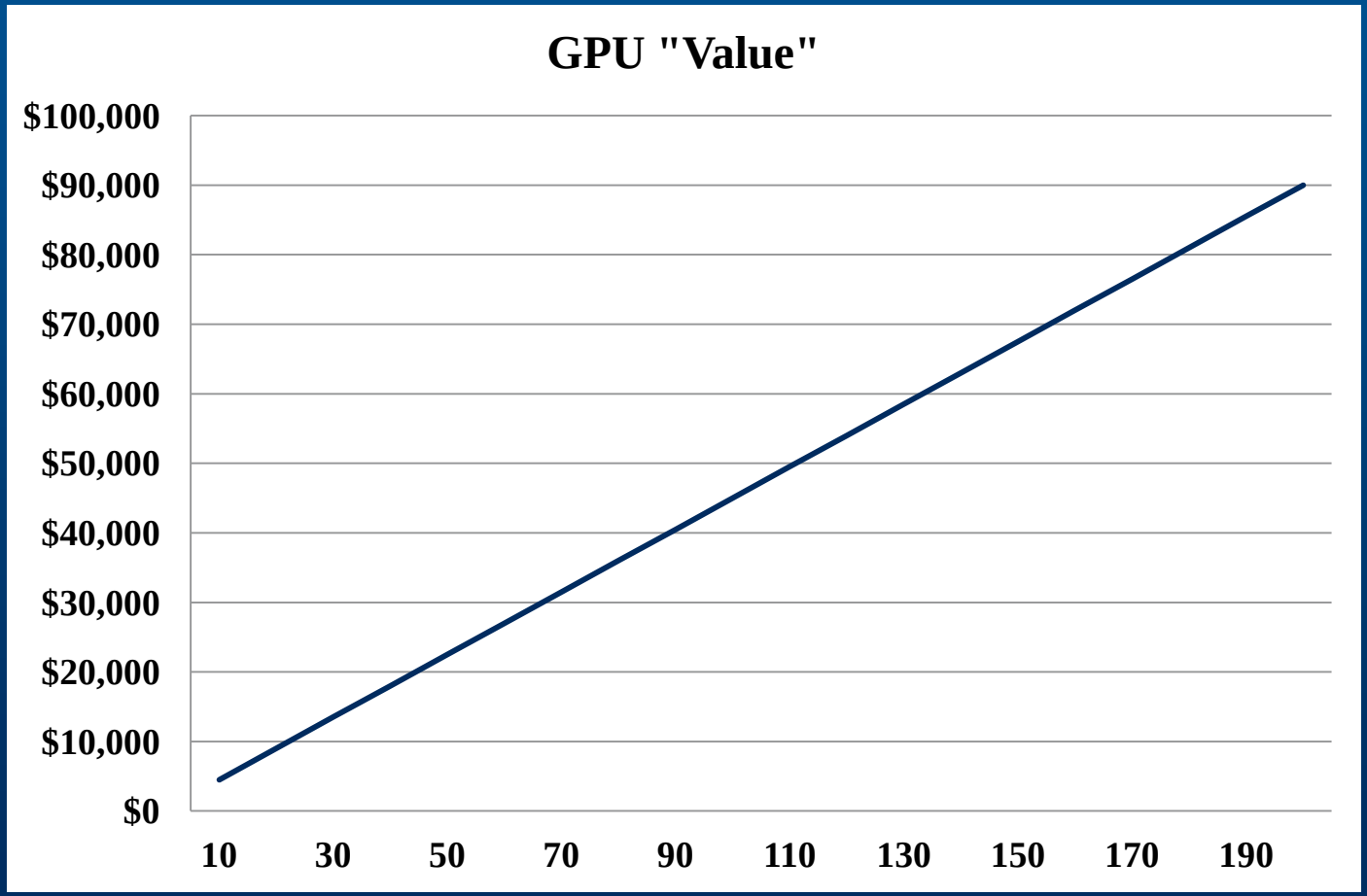
- Or $450 per core per year… if you want a DR site… $900

GPU “Value” For Our Model

A transparent breakdown

- C1060
  - 70x faster than CPU core
  - CPU costs $450 per core per year
  - 1 GPU yields $31,500 ($450 * 70)
- C2050
  - 140x faster than CPU core
  - CPU costs $450 per core per year
  - 1 GPU yields $63,000 ($450 * 140)

Generalized GPU “Value”



Overview

- Citadel's GPGPU Pilot Project
- Citadel's "Valuation" infrastructure
- Virtualized GPUs
- Demo

What is MDL?

- Application Accelerator
  - Users can significantly speed up applications by offloading the computationally expensive parts of their application. MDL enabled applications instantly have access to thousands of CPUs / GPUs without having to write any network communication code.

What is MDL?

- Application Accelerator
- Parallel Programming Tool
  - MDL gives your application thread like performance gains without having to write threaded code. Developers don't have to manage thread pools, mutex shared resources, or deal with reentrant code. MDL can provide users with access to 100s of thousands of concurrent threads.

What is MDL?

- Application Accelerator
- Parallel Programming Tool
- Grid Computing Ecosystem
  - MDL harnesses underutilized (as well as dedicated) computational resources across your organization providing centralized control and historical usage information.

What is MDL?

- Application Accelerator
- Parallel Programming Tool
- Grid Computing Ecosystem
- Supercomputer Infrastructure
  - MDL allows users to build very powerful grids of GPUs. A GPU grid can provide hundreds of TFLOPs of computational capacity at a fraction of the cost of a supercomputer.

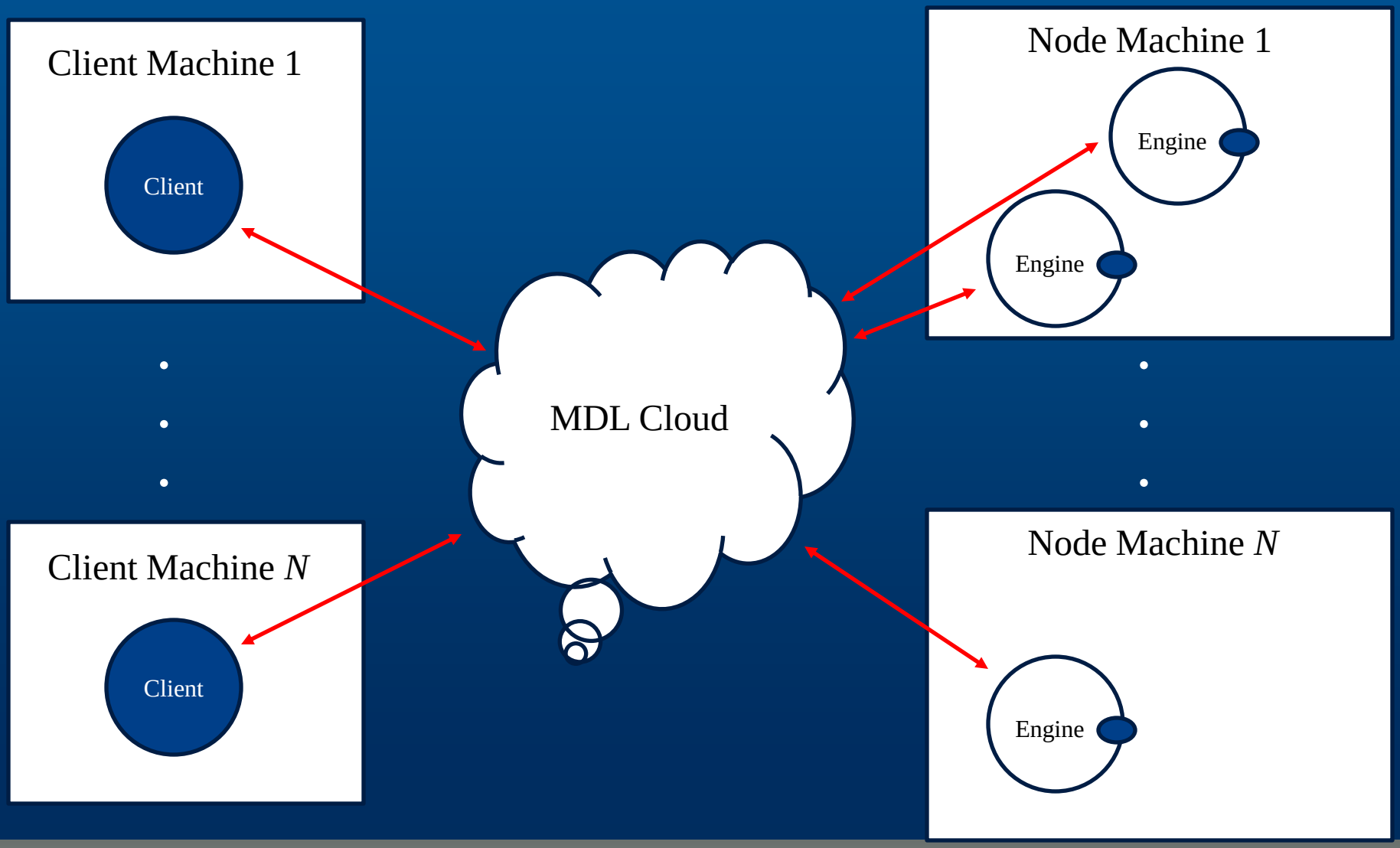
What is MDL?

- Application Accelerator
- Parallel Programming Tool
- Grid Computing Ecosystem
- Supercomputer Infrastructure
- Green
  - By harnessing a companies underutilized computational resources and enabling large scale GPGPU computing MDL reduces power consumption significantly.

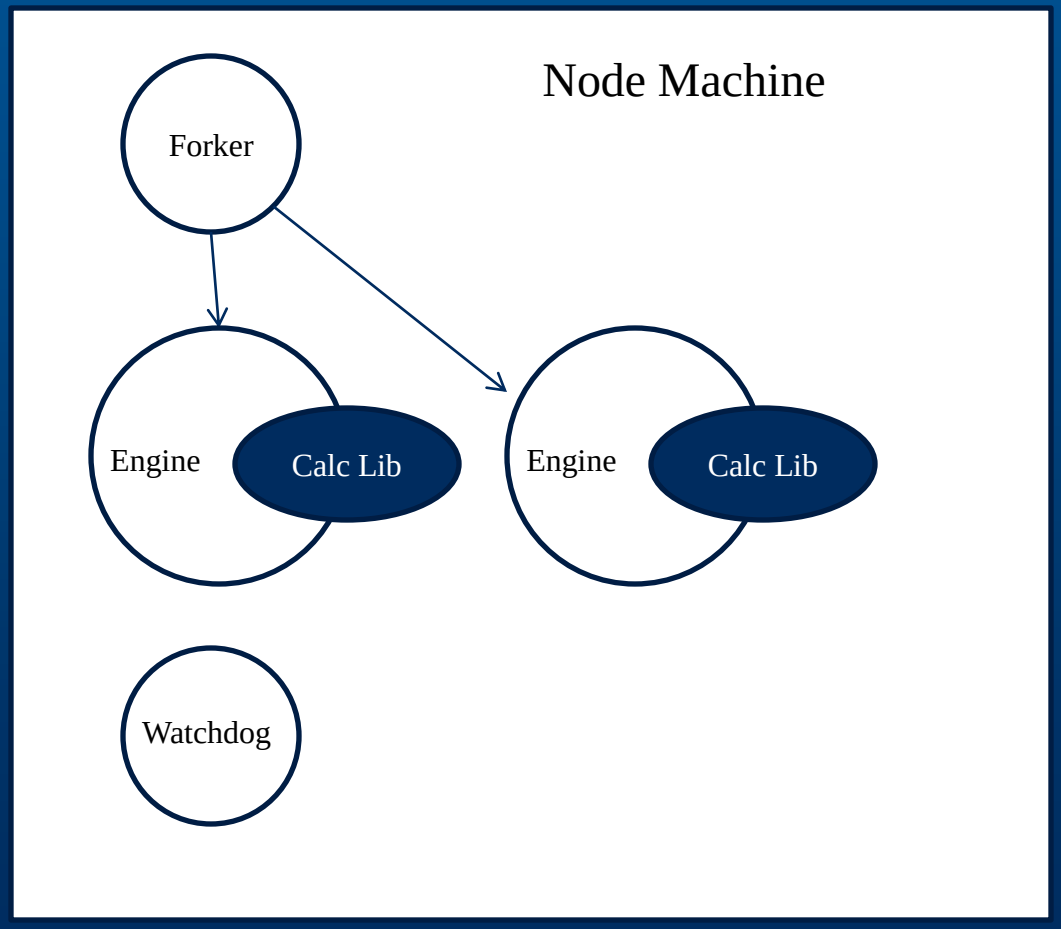
What is MDL?

- Application Accelerator
- Parallel Programming Tool
- Grid Computing Ecosystem
- Supercomputer Infrastructure
- Green
- Saves Money
  - MDL saves money on hardware, power, data center space, and software development.

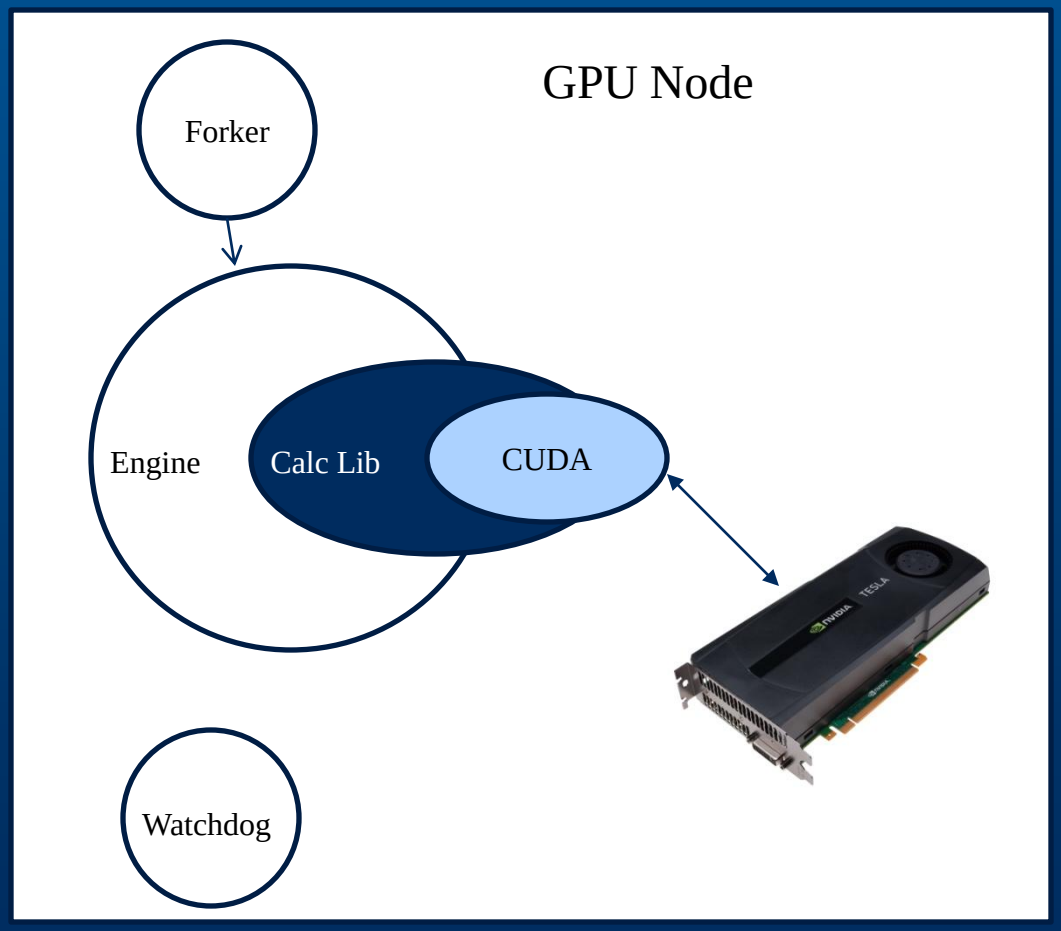
How MDL Works



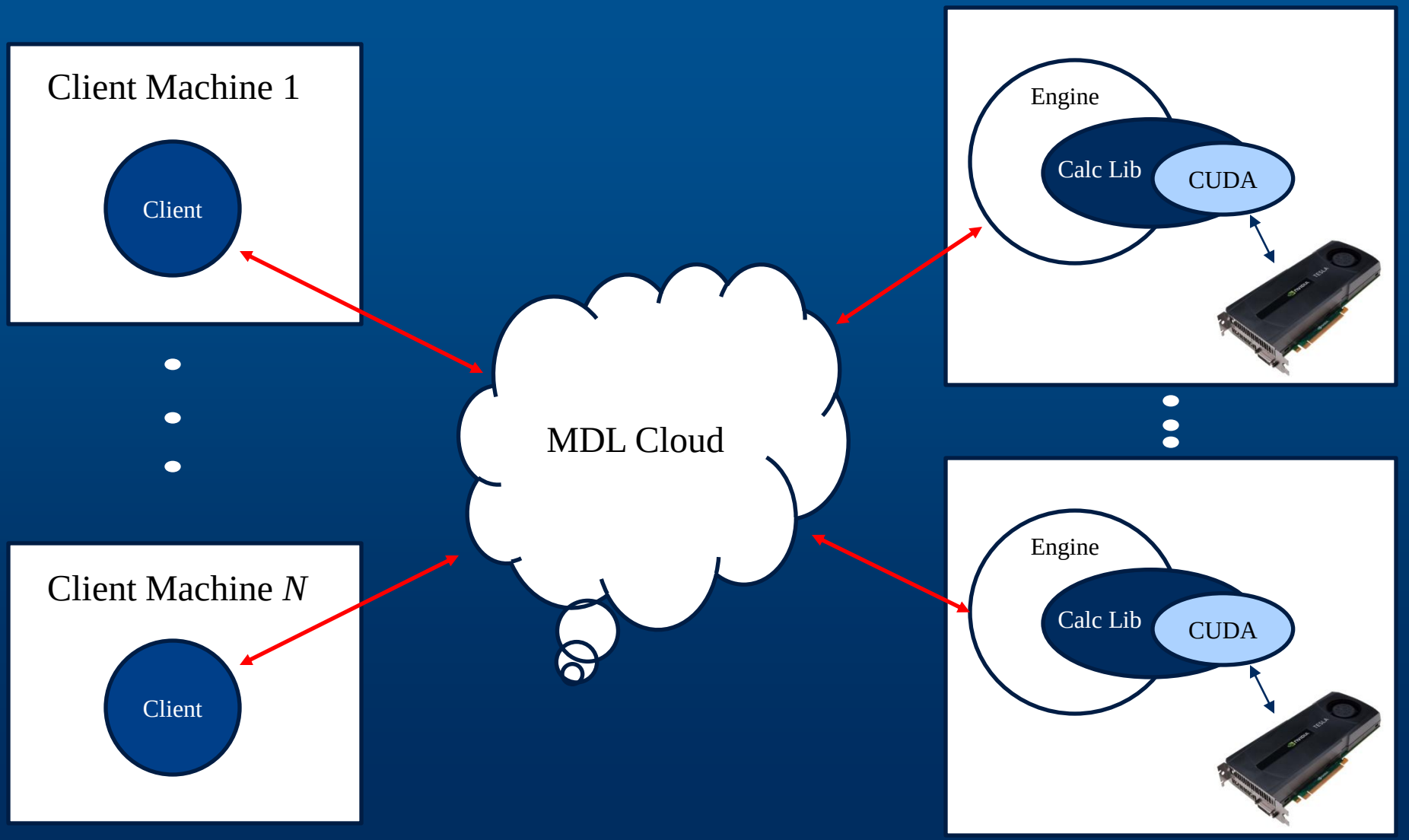
How MDL Works



GPGPU and MDL Integration



GPGPU and MDL Integration



Overview

- Citadel's GPGPU Pilot Project
- Citadel's "Valuation" infrastructure
- Virtualized GPUs
- Demo

Types of Compute Nodes

- Dedicated CPU Server Nodes
- Scavenged CPU Server Nodes
- Dedicated GPU Server Nodes
- Scavenged Desktop Nodes…?

Scavenging Unutilized Desktop Resources

» Most modern desktop systems are significantly underutilized

Scavenging Unutilized Desktop Resources

» Most modern desktop systems are significantly underutilized

» By installing a VM on every desktop we can “stitch together” underutilized computational resources from each desktop and provide centralized access and control of those resources with MDL.

Scavenging Unutilized Desktop Resources

» Most modern desktop systems are significantly underutilized

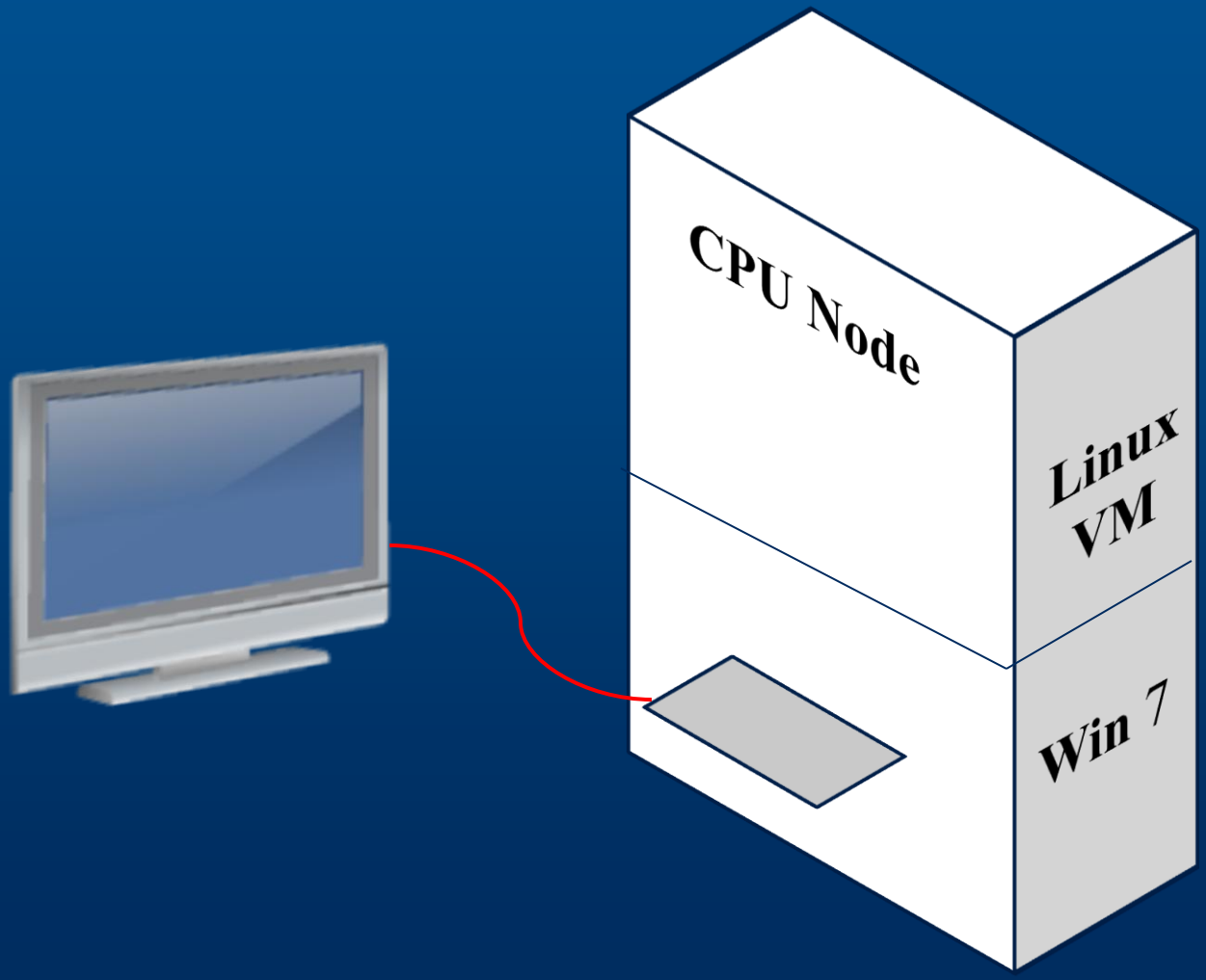
» By installing a VM on every desktop we can “stitch together” underutilized computational resources from each desktop and provide centralized access and control of those resources with MDL.

» The VM provides a layer of insulation between the calculations running on the desktop and the desktop owners processes.

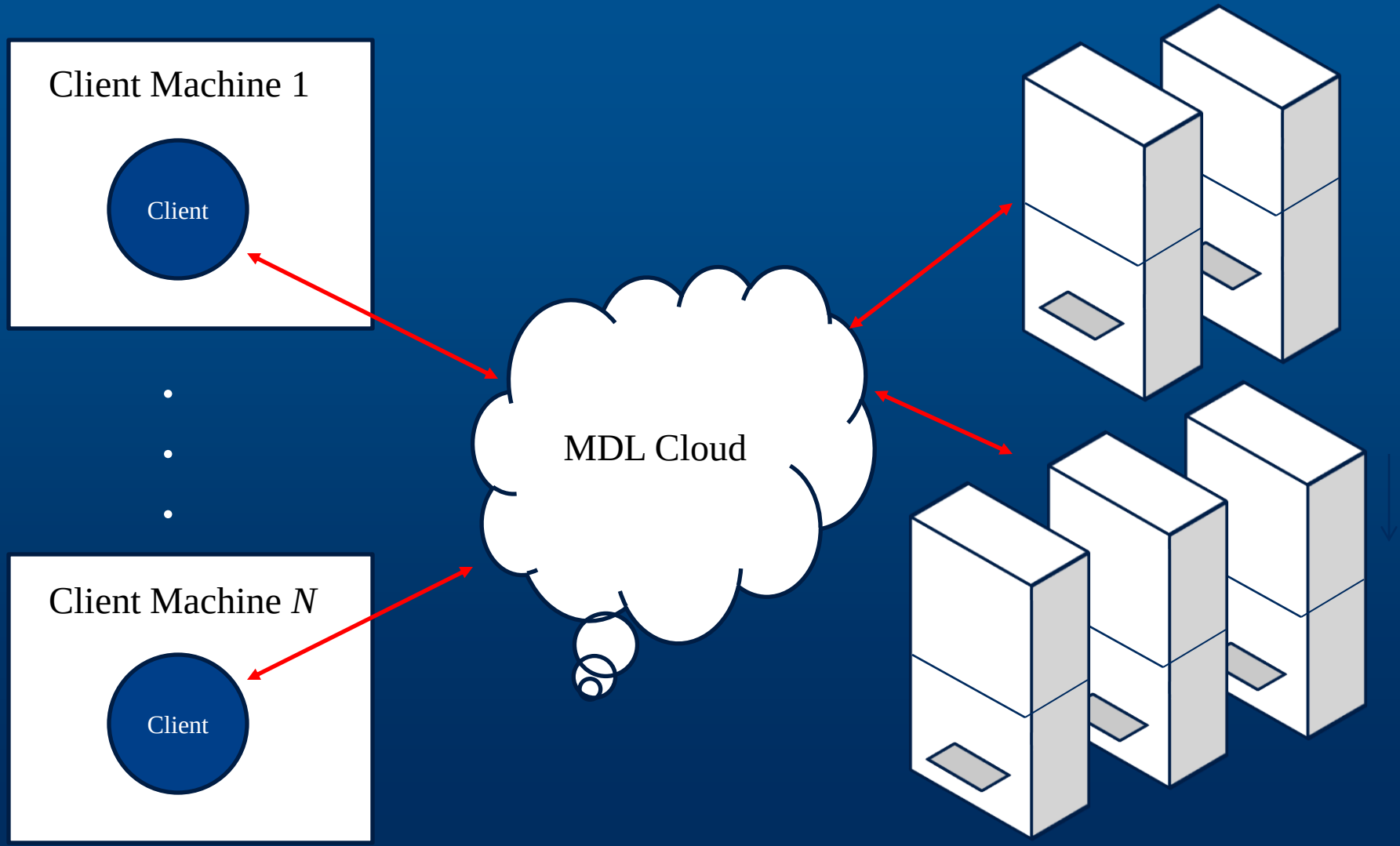
Scavenging Unutilized Desktop Resources

- Most modern desktop systems are significantly underutilized
- By installing a VM on every desktop we can “stitch together” underutilized computational resources from each desktop and provide centralized access and control of those resources with MDL.
- The VM provides a layer of insulation between the calculations running on the desktop and the desktop owners processes.
- The VM allows us to use Windows based desktops to perform Linux based calculations.

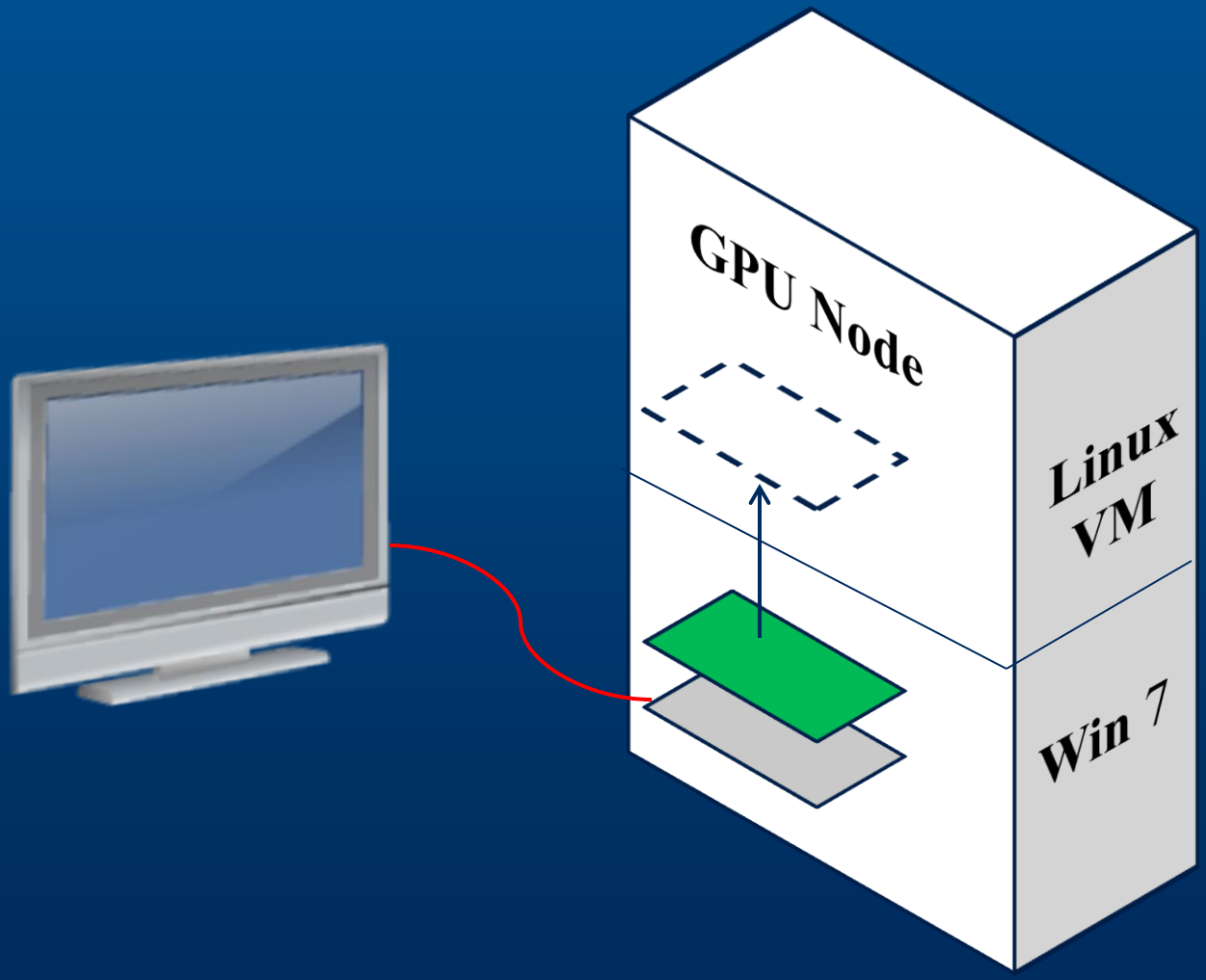
Virtualized Desktop CPU Nodes



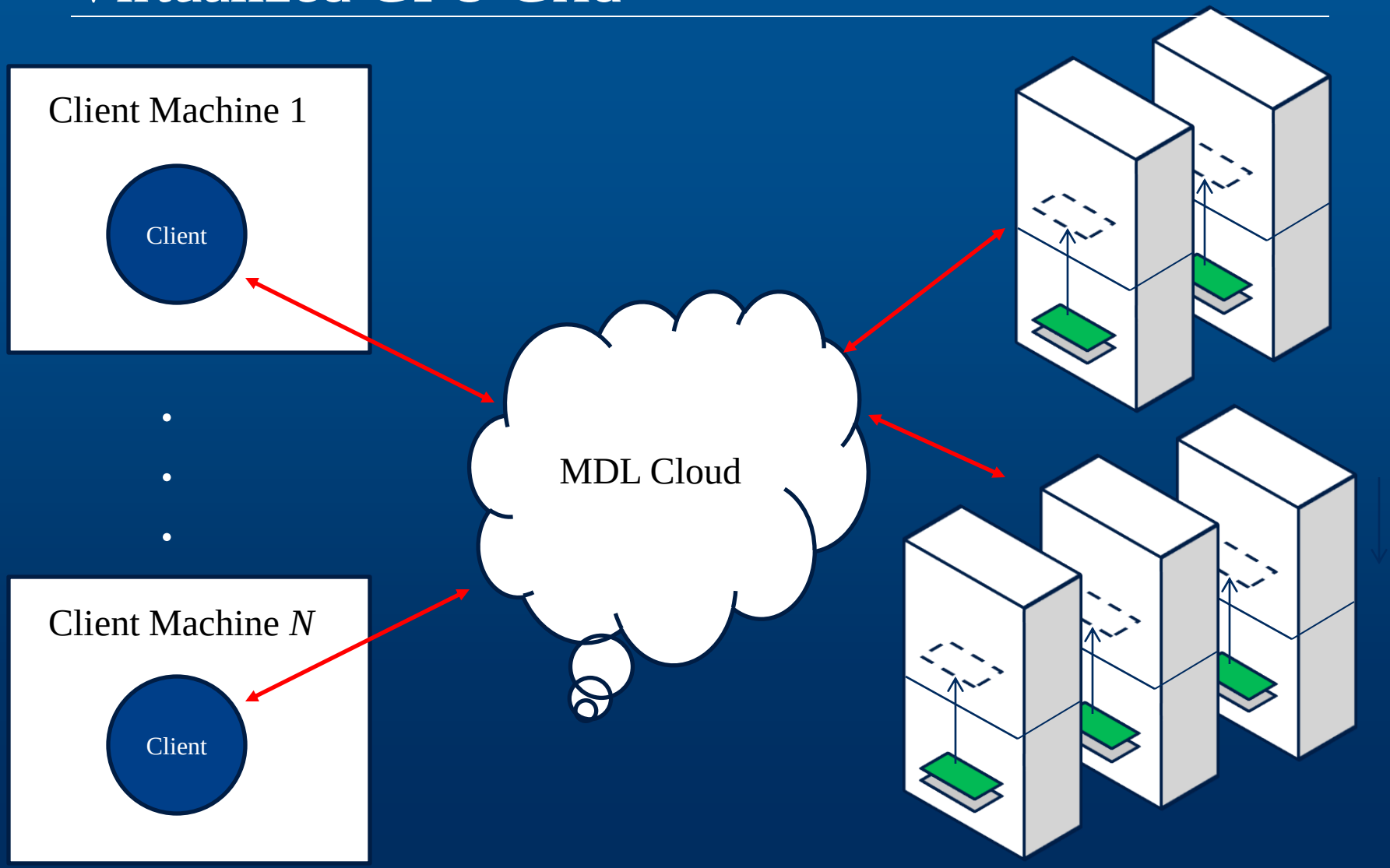
Virtualized Desktop CPU Grid



Virtualized GPU Nodes



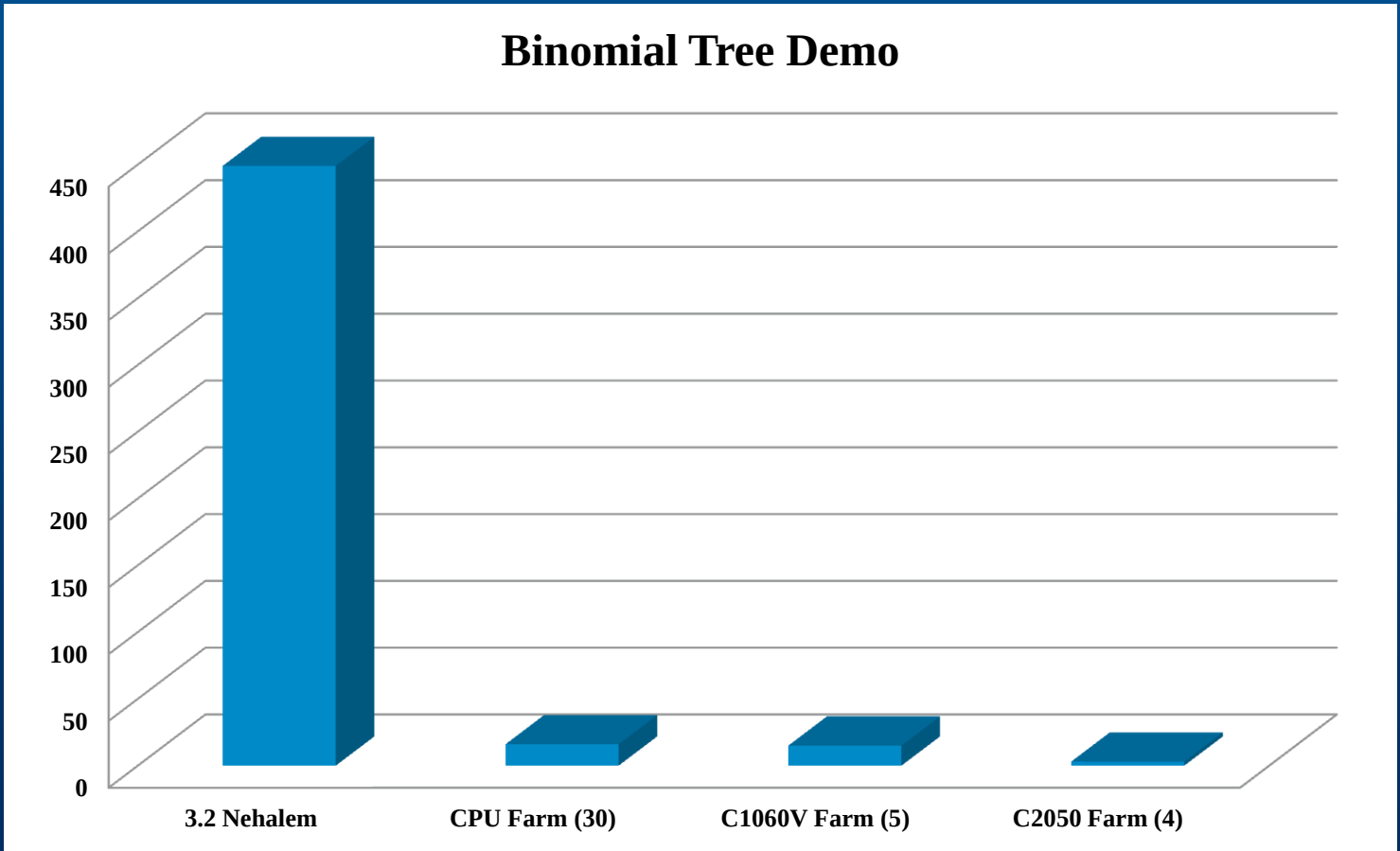
Virtualized GPU Grid



Overview

- Citadel's GPGPU Pilot Project
- Citadel's "Valuation" infrastructure
- Virtualized GPUs
- Demo

Supercomputer Demo



Why call it a “Supercomputer”?

- 200 desktops utilized in this manner will provide you with 100 TFLOPS of computational capacity at a cost of less than $1,000,000

Why call it a “Supercomputer”?

- 200 desktops utilized in this manner will provide you with 100 TFLOPS of computational capacity at a cost of less than $1,000,000

- As of June 2010 100 TFLOPS would land you somewhere around position 70 of the top 500 supercomputers in the world

Questions

- We are currently working with OEMs to turn our internal technologies into external products
- More information about our products can be found at http://www.citadeltechnology.com
- You can contact me directly scott.donovan@citadeltechnology.com