Integrated GPUs for Green High-Performance Computing

Tom Scogland
Heshan Lin
Wu-chun Feng

SyNeRG
http://synergy.cs.vt.edu

Virginia Tech
Invent the Future

GPUs

Powerful

http://synergy.cs.vt.edu

Invent the Future

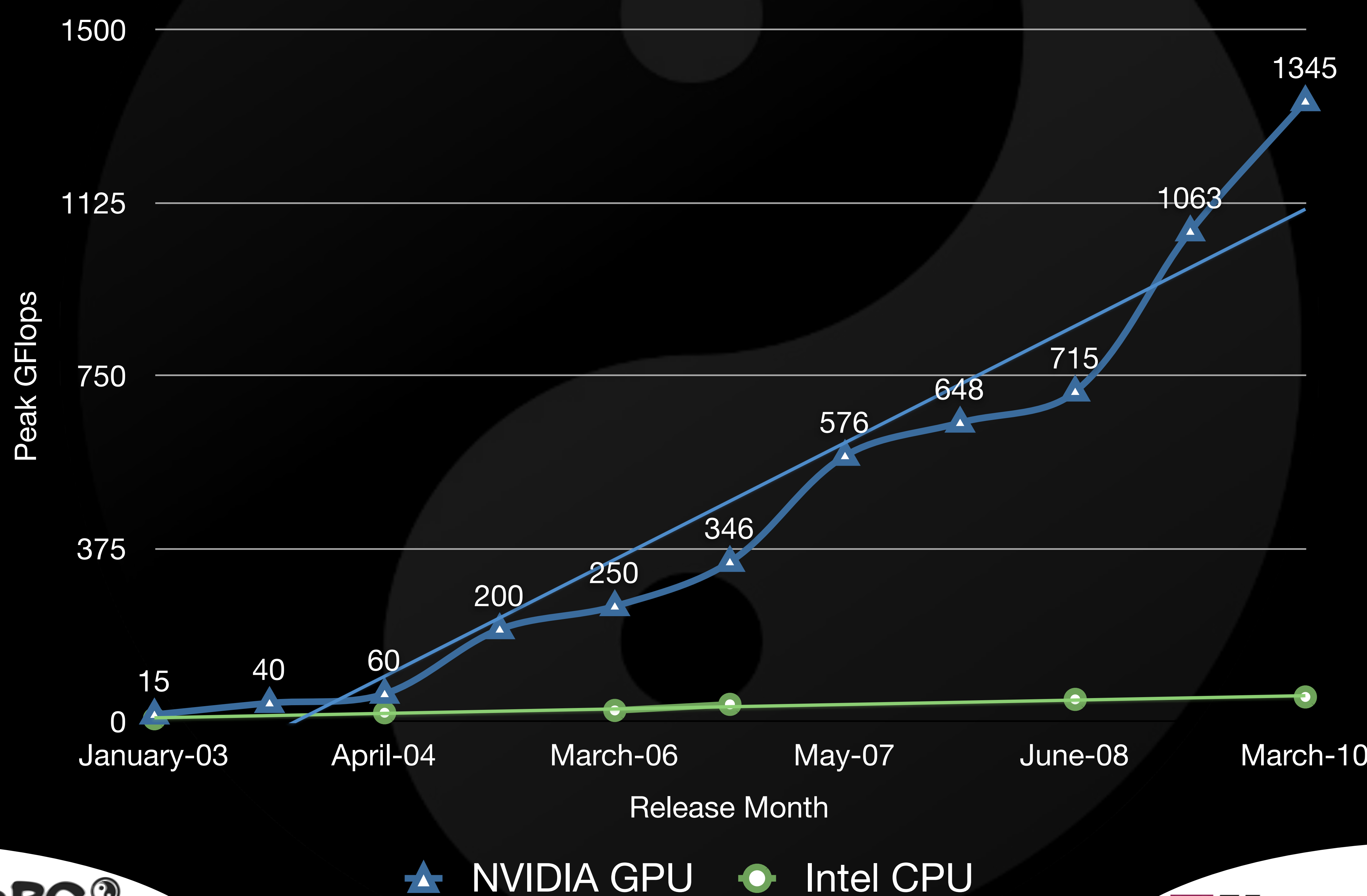
Powerful: Graphics

Powerful: Graphics

Powerful: Graphics

Powerful: Computing

Powerful: Computing



SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

Powerful: Computing

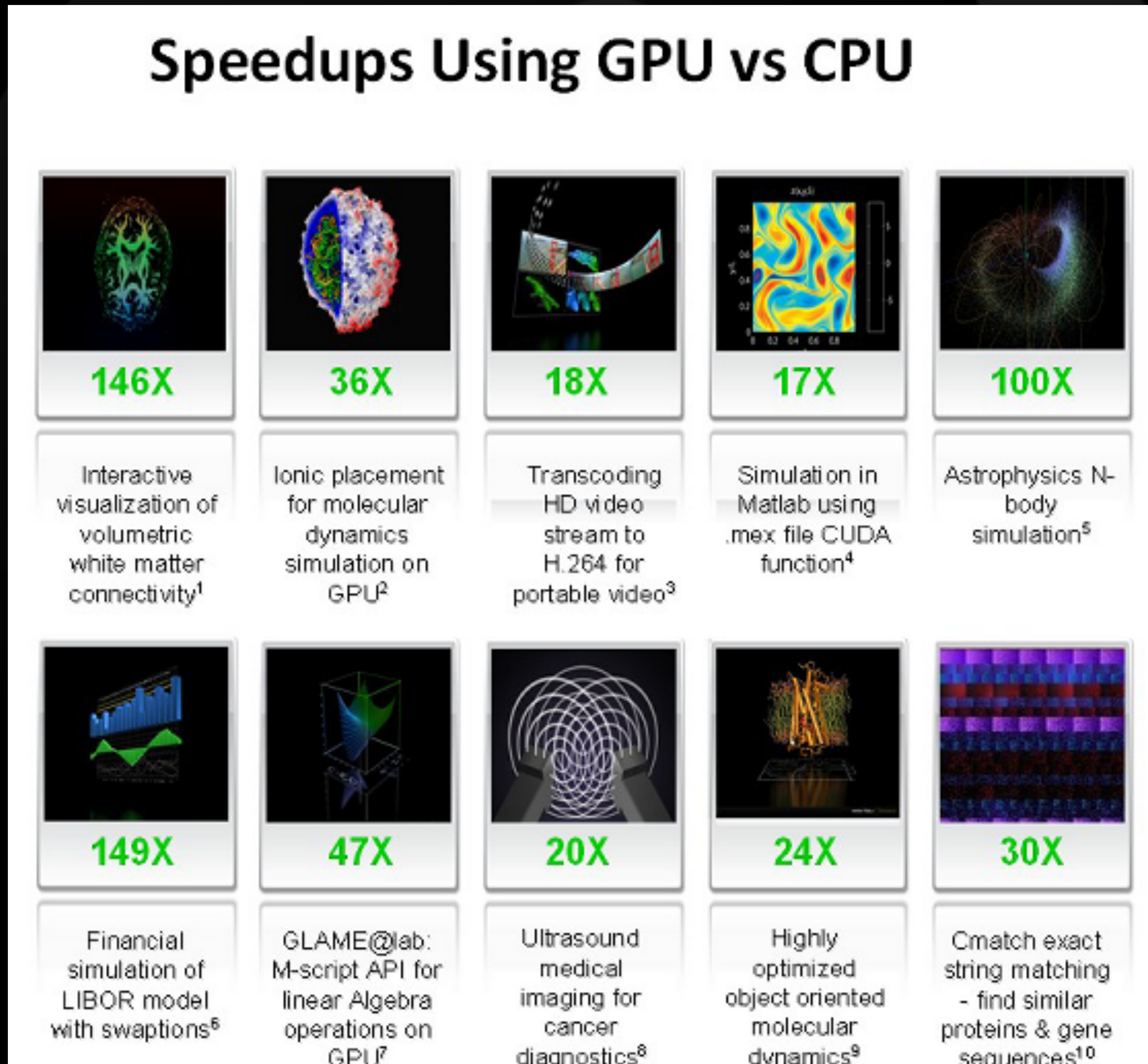
NVIDIA®
CUDA®

Powerful: Computing

NVIDIA
CUDA

OpenCL

Powerful: Computing



Powerful: Computing

TOP500® SUPERCOMPUTER SITES

Rank	Site	Computer
1	Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron Six Core 2.6 GHz Cray Inc.
2	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU Dawning
3	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM
4	National Institute for Computational Sciences/University of Tennessee United States	Kraken XT5 - Cray XT5-HE Opteron Six Core 2.6 GHz Cray Inc.

Powerful: Computing

TOP500 SUPERCOMPUTER SITES

Rank	Site	Computer
1	Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron Six Core 2.6 GHz Cray Inc.
2	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU Dawning
3	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM
4	National Institute for Computational Sciences/University of Tennessee United States	Kraken XT5 - Cray XT5-HE Opteron Six Core 2.6 GHz Cray Inc.

Green

Green?

http://synergy.cs.vt.edu

What is Green?

Green

Green

Energy Efficiency

http://synergy.cs.vt.edu

Invent the Future

Green

Energy Efficiency

Power Efficiency

Green: Energy Efficiency

THE GREEN 500™

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
1	773.38	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
1	773.38	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
4	492.64	National Supercomputing Centre in Shenzhen (NSCS)	Dawning Nebulae, TC3600 blade CB60-G2 cluster, Intel Xeon 5650/ nVidia C2050, Infiniband	2580

Green: Energy Efficiency

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
1	773.38	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
1	773.38	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
4	492.64	National Supercomputing Centre in Shenzhen (NSCS)	Dawning Nebulae, TC3600 blade CB60-G2 cluster, Intel Xeon 5650/ nVidia C2050, Infiniband	2580

Related Work

SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

Related Work

- Huang S, Xiao S, Feng W (2009) **"On the Energy Efficiency of Graphics Processing Units for Scientific Computing**." In Proceedings of the 5th IEEE Workshop on High-Performance, Power-Aware Computing (in conjunction with the 23rd International Parallel & Distributed Processing Symposium), Rome, Italy, May 2009.

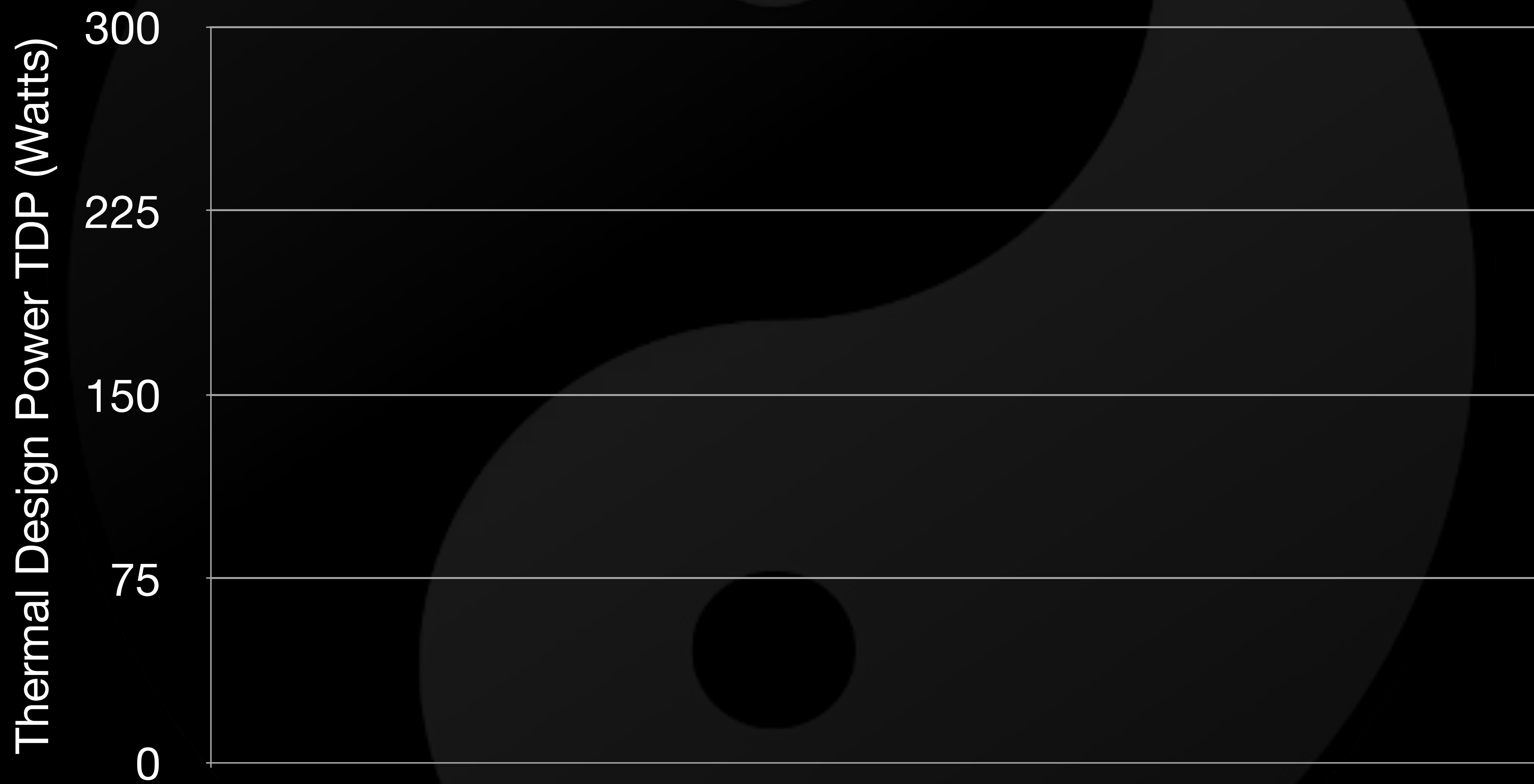
Related Work

- Huang S, Xiao S, Feng W (2009) **“On the Energy Efficiency of Graphics Processing Units for Scientific Computing**.“ In Proceedings of the 5th IEEE Workshop on High-Performance, Power-Aware Computing (in conjunction with the 23rd International Parallel & Distributed Processing Symposium), Rome, Italy, May 2009.
- Rofouei M, Stathopoulos T, Ryffel S, Kaiser W, Sarrafzadeh M (2008) “**Energy-Aware High Performance Computing with Graphic Processing Units.**” In: Workshop on Power Aware Computing and System

Green: Power

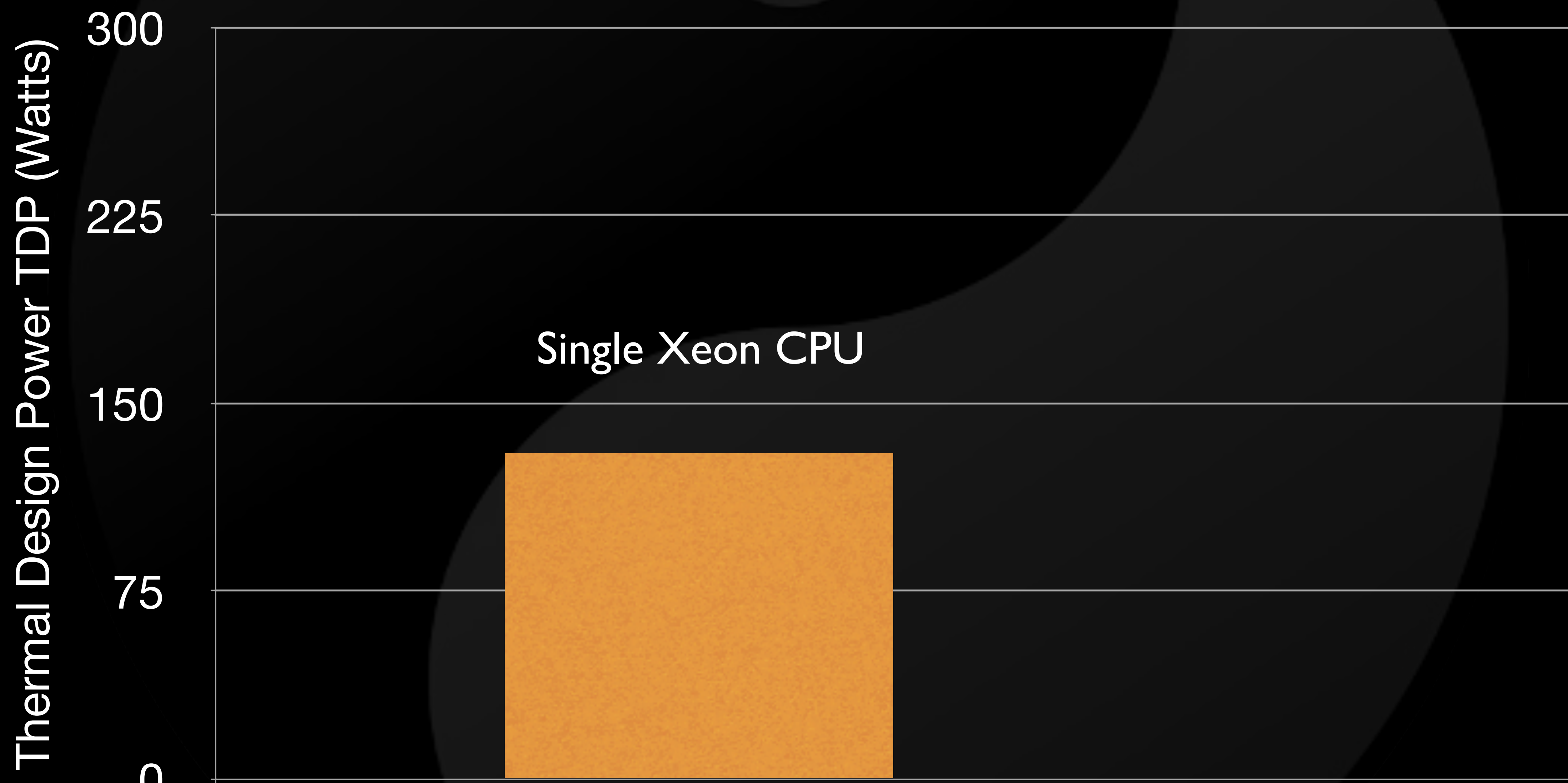
Power of processing unit only

Green: Power



Power of processing unit only

Green: Power

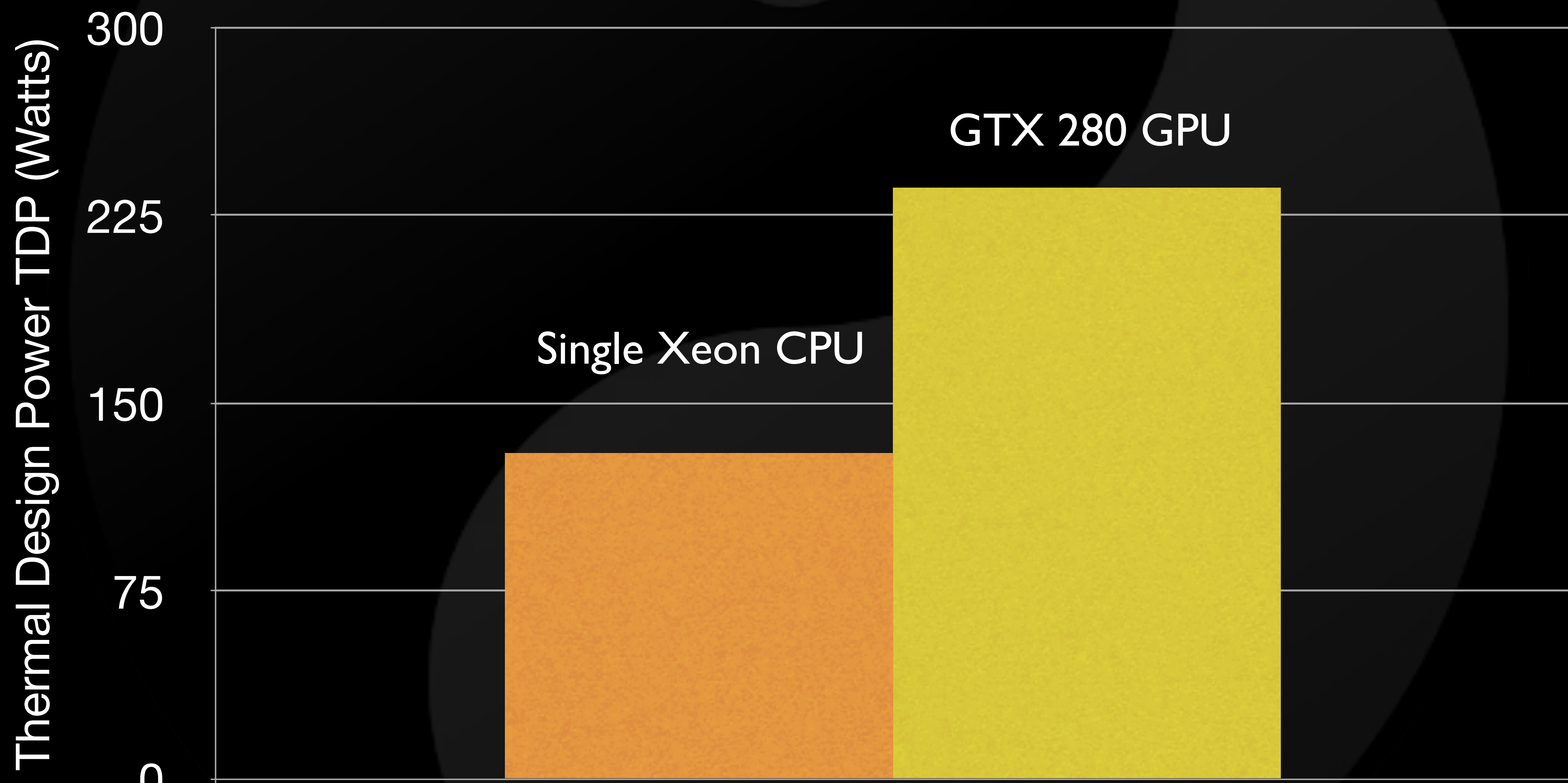


Power of processing unit only

SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

Green: Power



Power of processing unit only

Green: Power Efficiency

- Systems which are
  - Embedded
  - Portable
  - Power constrained
  - Cooling constrained

SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

What We Need

What We Need

?

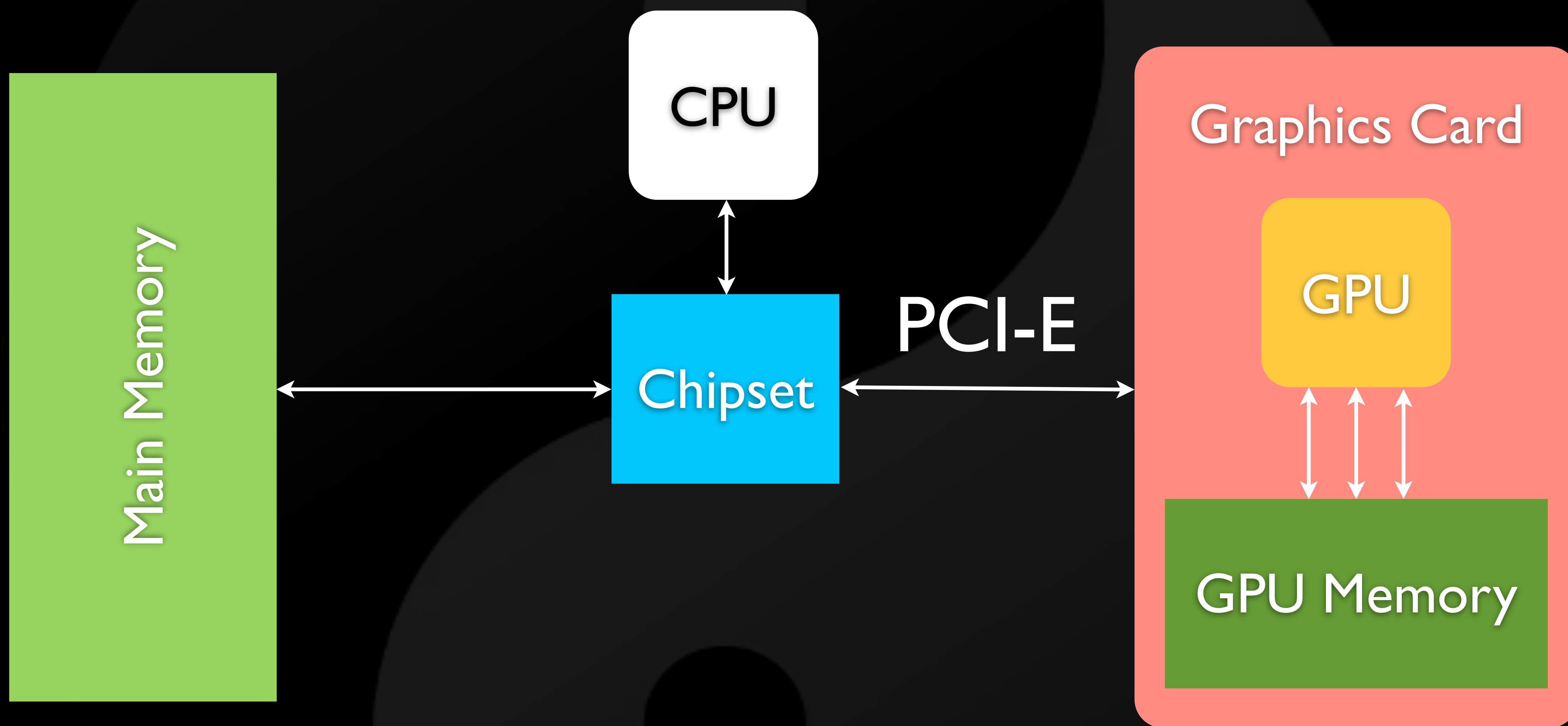
Integrated GPUs

Integrated GPUs: What?

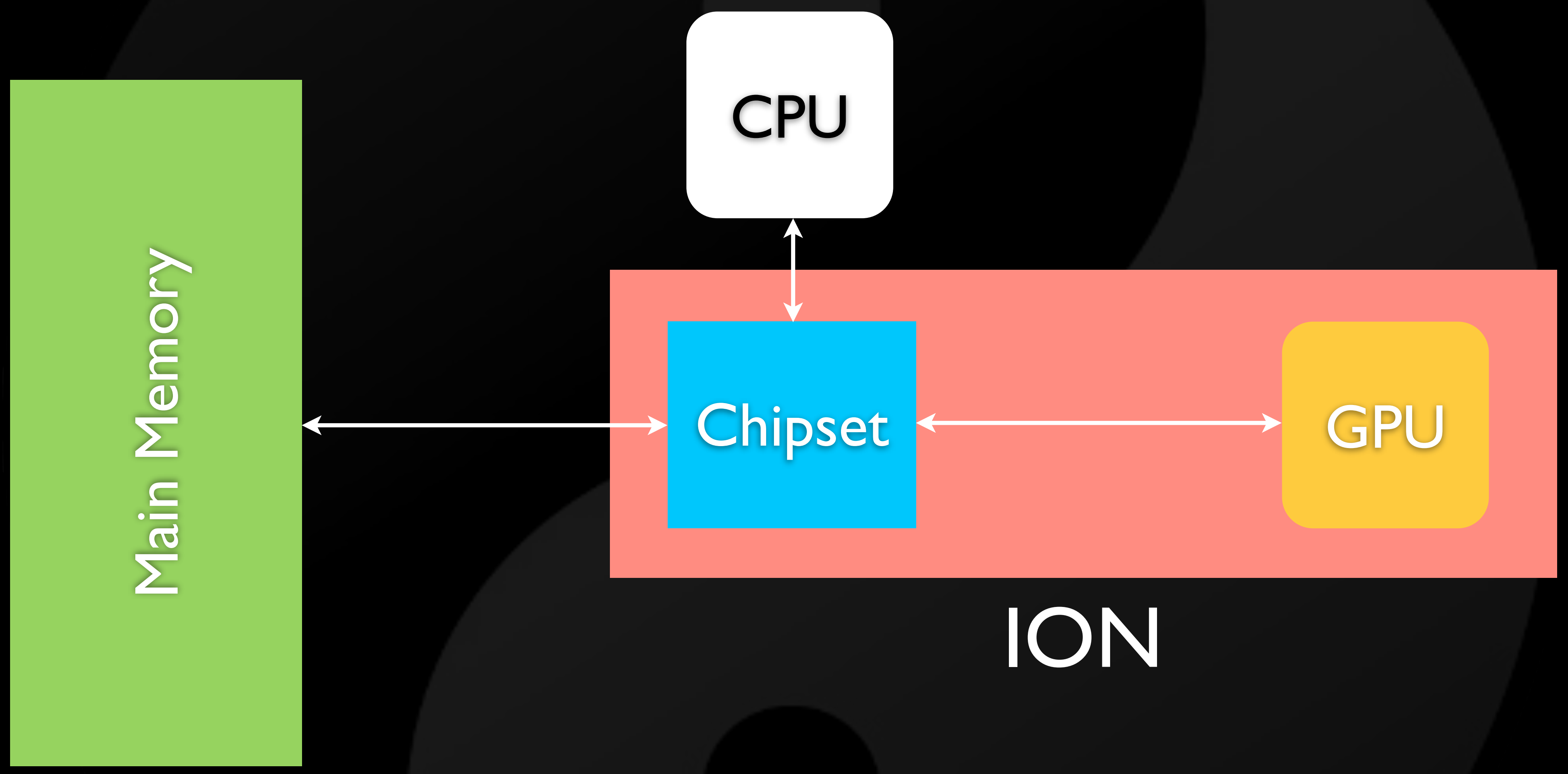
Discrete GPU

Discrete GPU

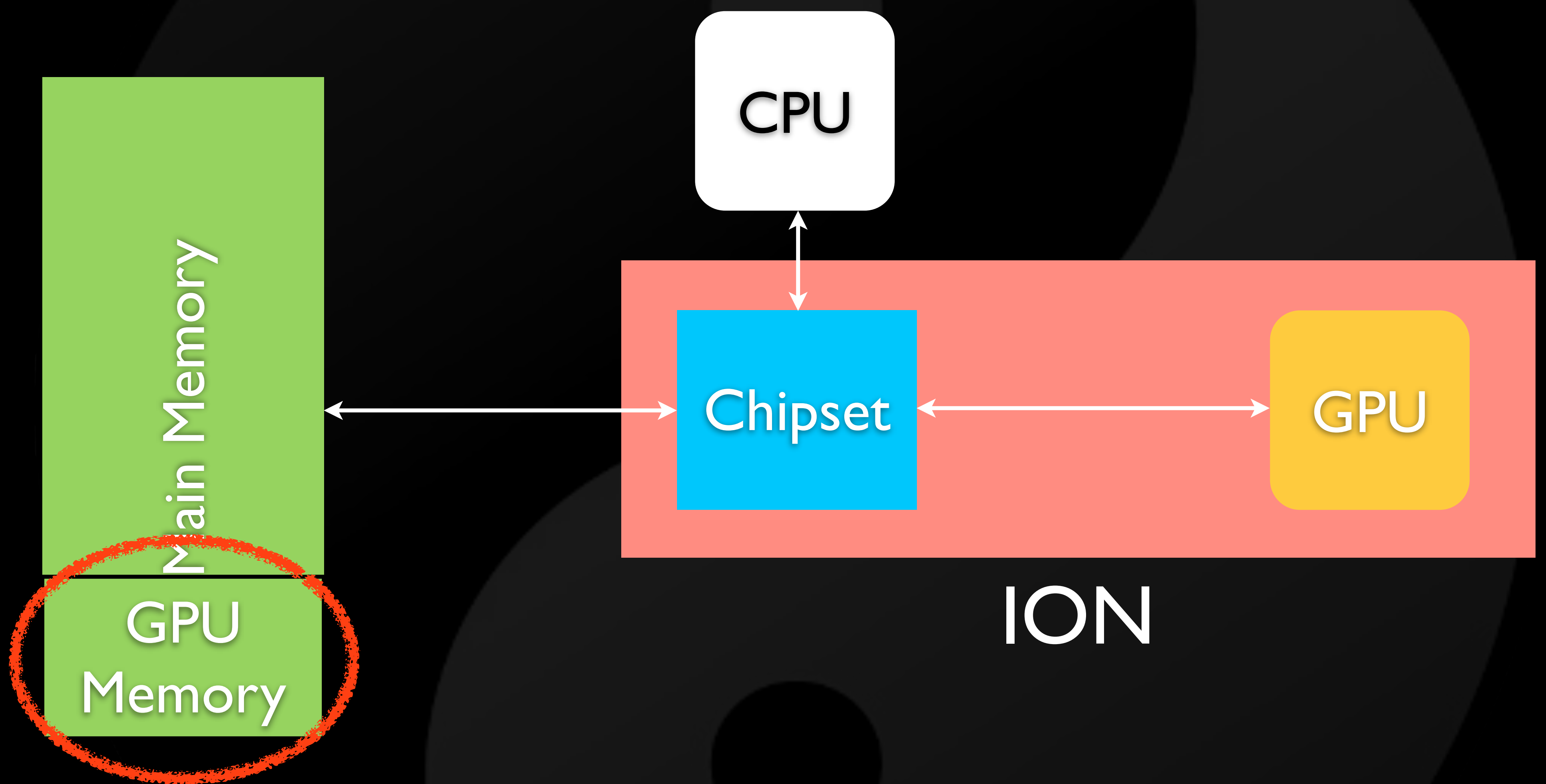
Discrete GPU



Integrated GPU



Integrated GPU



Integrated GPUs

Common Perception

Integrated GPUs

Discrete GPU =

Integrated GPU =

Integrated GPUs

Discrete GPU = Fast

Integrated GPU =

Integrated GPUs

Discrete GPU = Fast

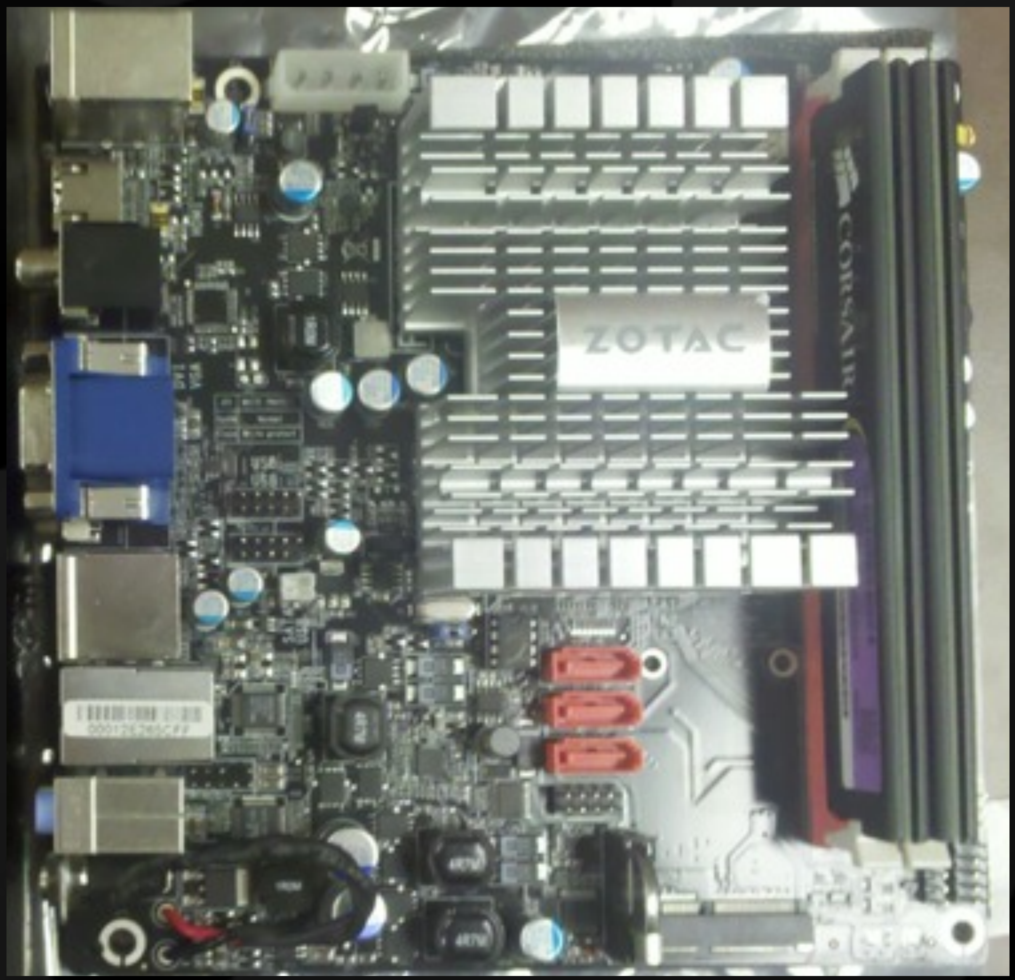
Integrated GPU = Slow

Integrated GPUs

Discrete GPU = Fast

Integrated GPU = Slow?

Tested GPUs



Tested GPUs

	NVIDIA GTX 280	NVIDIA ION (GeForce 9400)
Multiprocessors	30	2
Cores	240	16
Clock Rate	1.3 GHz	1.1 GHz
Memory	1GB (device)	256MB (system)

Tested Systems

	GTX 280	ION
CPU	Xeon e5405 x2	Atom 230
Cores	8	2
Clock Rate	2 GHz	1.6 GHz
Memory	4GB	3GB

Software

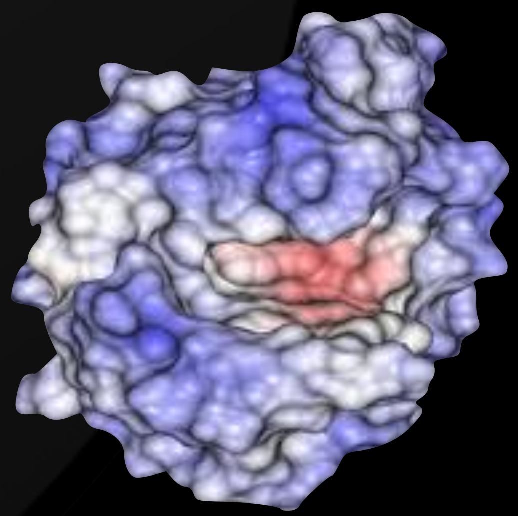
- OS: Linux
- Distribution: Ubuntu Hardy
- CUDA SDK 3.0
- GCC 4.2: all optimizations enabled

Tested Applications

- Tested 4 dwarfs
  - N-Body
  - Structured grid
  - Dynamic programming
  - Dense linear algebra

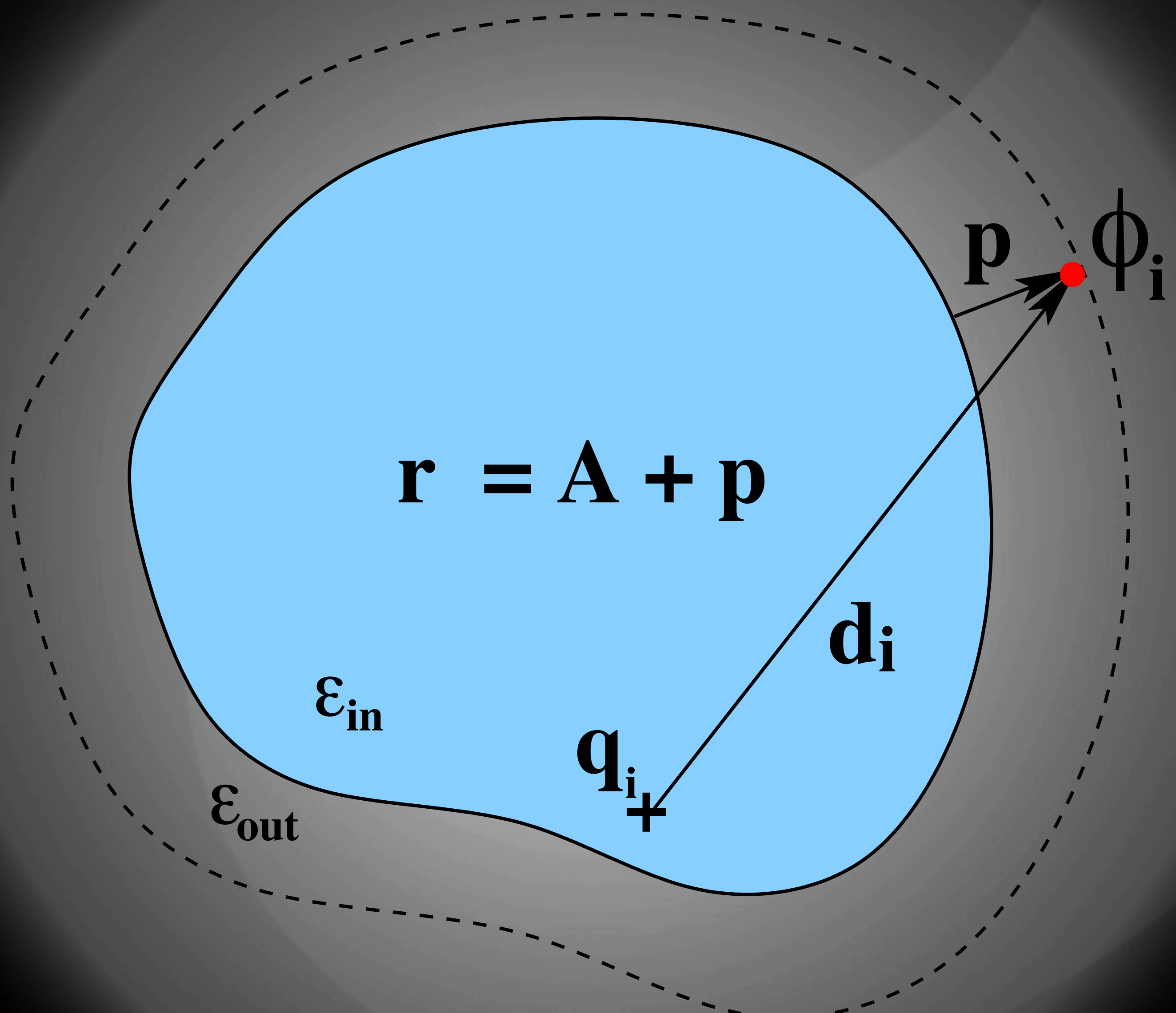
N-Body: GEM

- Application
  - Molecular dynamics
  - Computes electrostatic potential along the surface of macromolecules
- Signature
  - Compute intensive



GEM: Signature

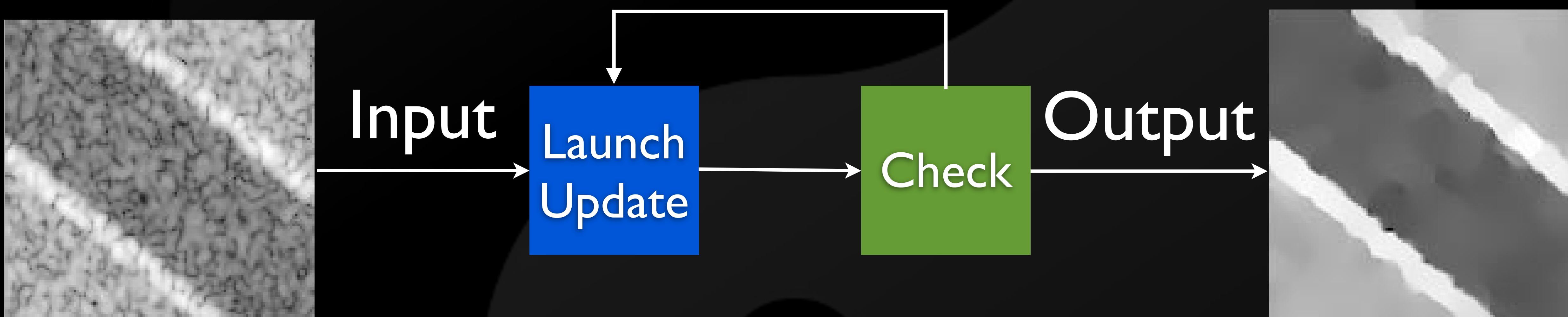
All-Pairs Vertices (ϕi) <-> Atoms (qi)



Structured Grid: SRAD

- Application
  - Image cleanup
  - Used to smooth sonic and radar imagery without losing important detail
- Signature
  - Kernel launch intensive

SRAD: Signature



Dynamic Programming: Needleman Wunsch (NW)

- Application
  - Sequence alignment
  - NW performs a global alignment between two sequences, usually DNA
- Signature
  - Synchronization intensive

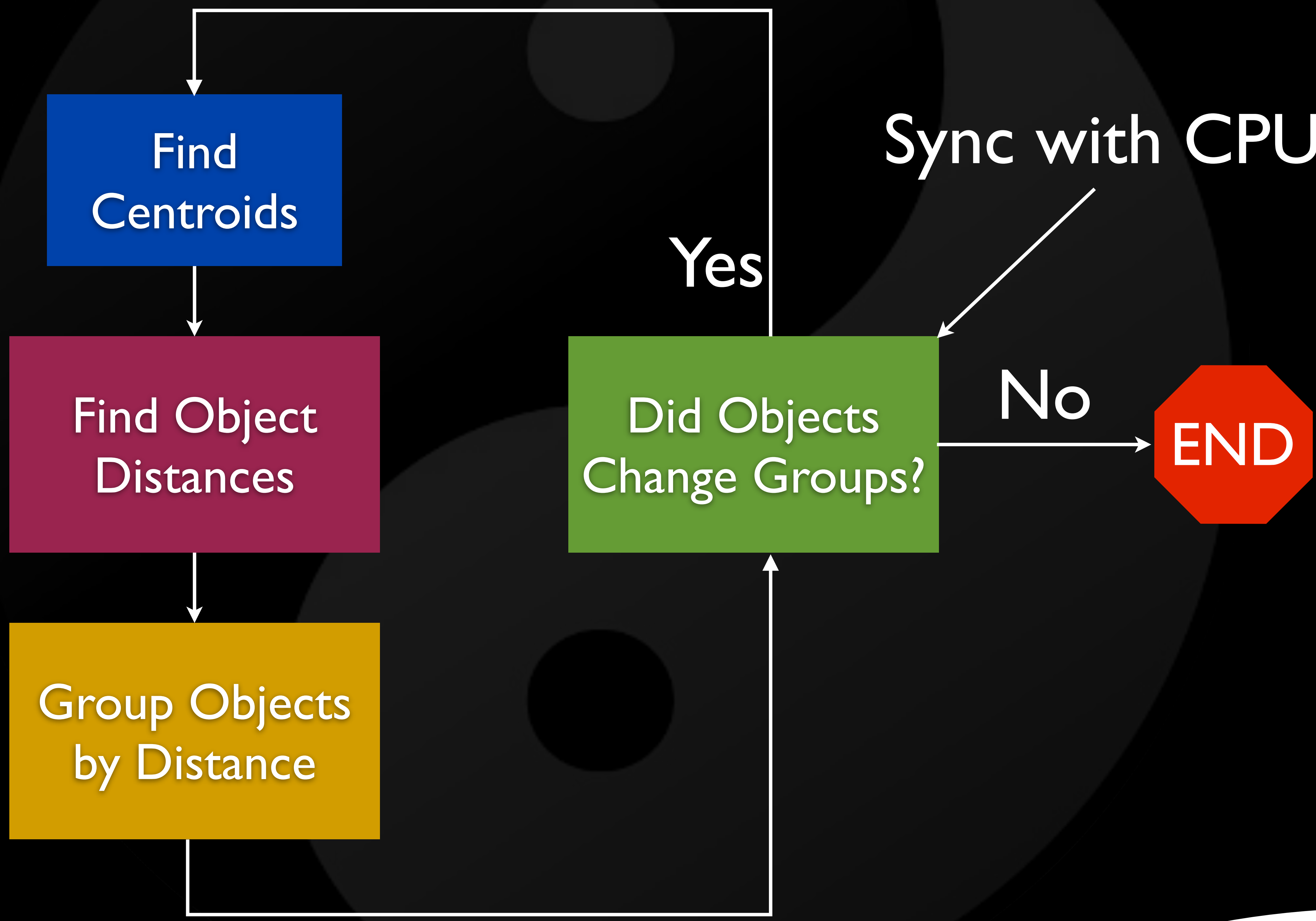
NW: Signature

	G	A	A	T	T	C	A
G	1	1	1	1	1		
G	1	1	1	1			
A	1	1	2				
T	1	2					
C	1						
G							

Dense Linear Algebra: K-Means

- Application
  - Clustering
  - Uses iterative refinement to cluster items in a space
- Signature
  - PCI-E Memory transfer intensive

K-Means: Signature



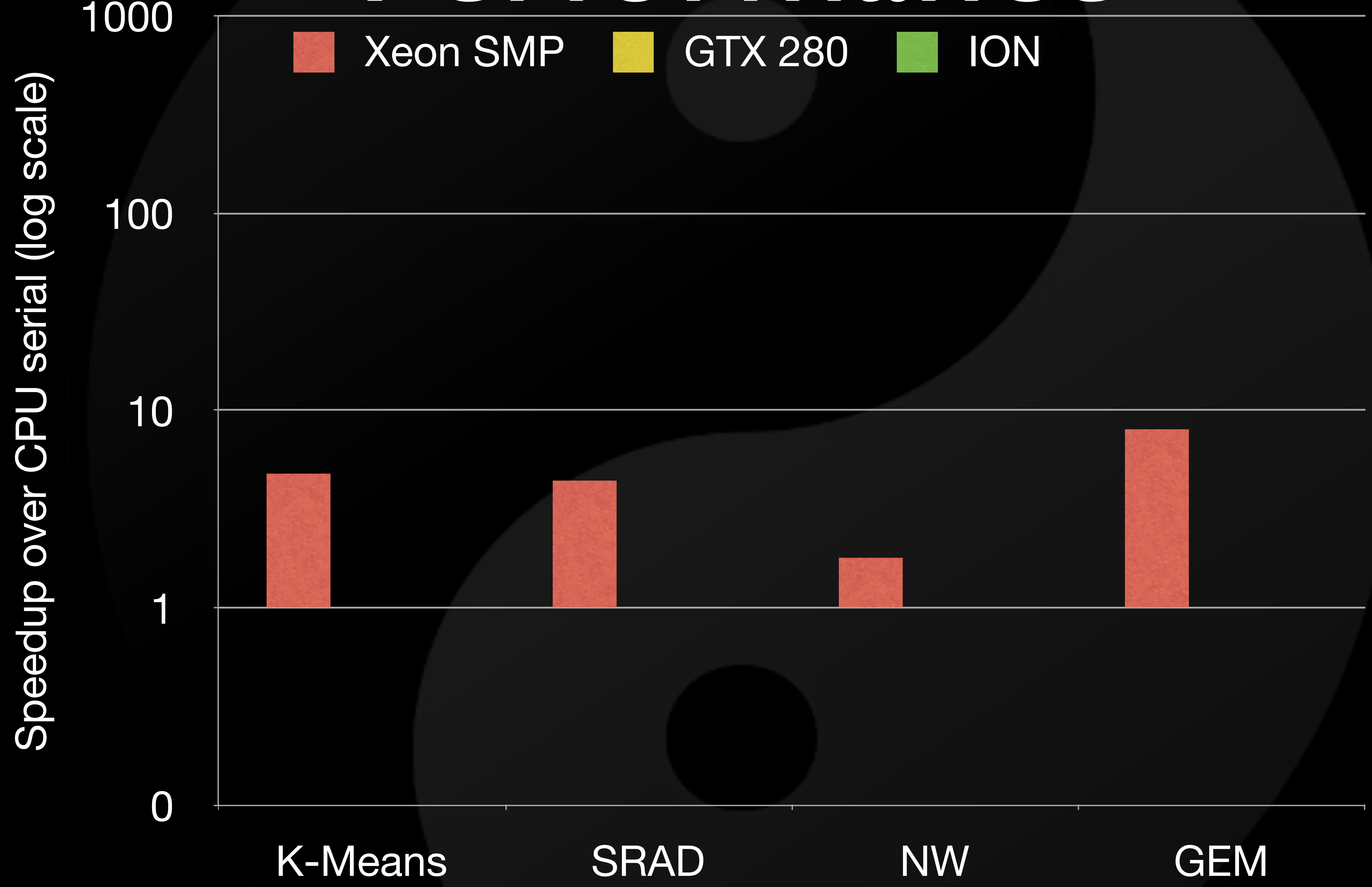
Results

Performance

Performance

■ Xeon SMP ■ GTX 280 ■ ION

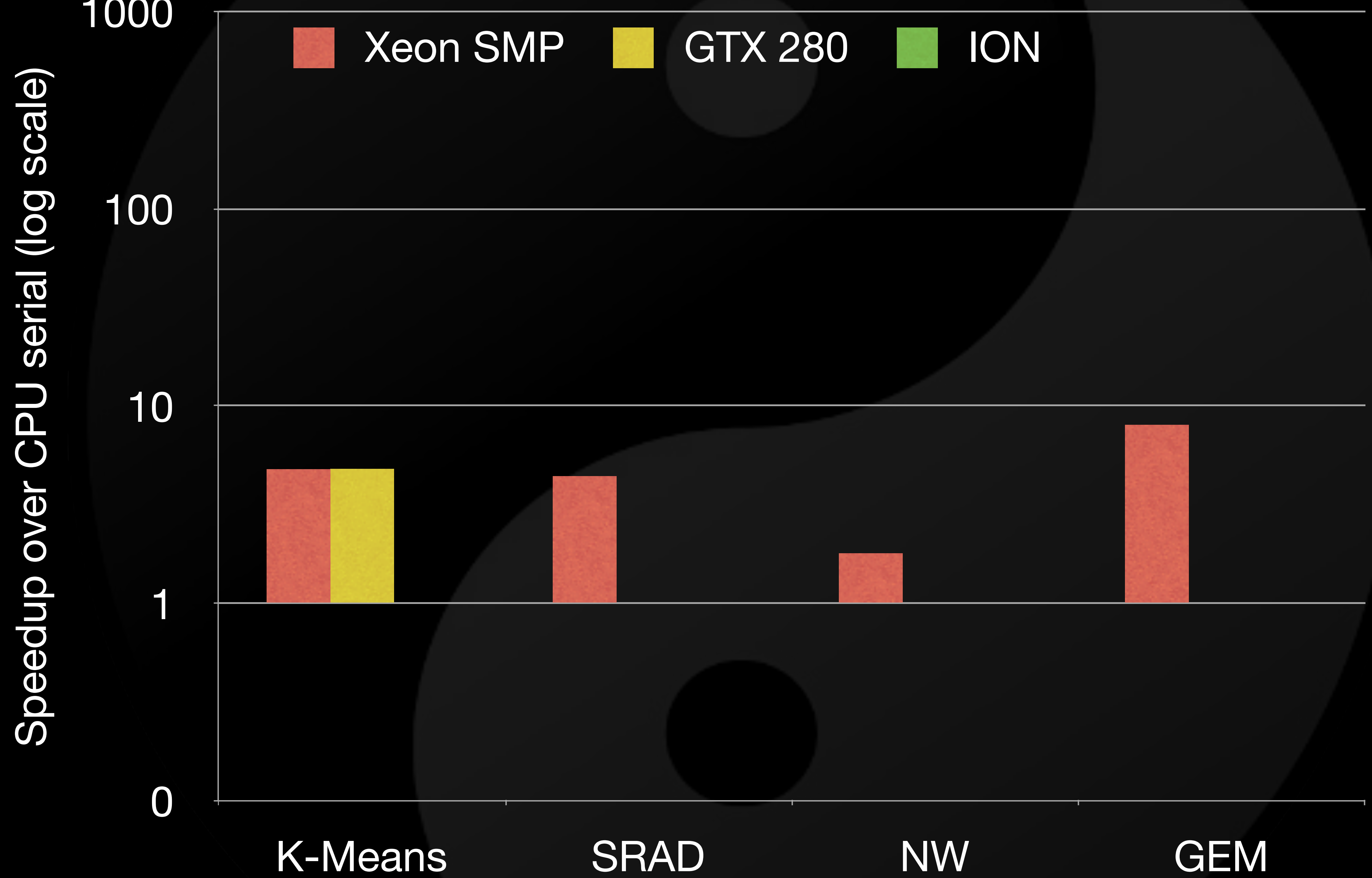
Performance



SyNeRGy
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

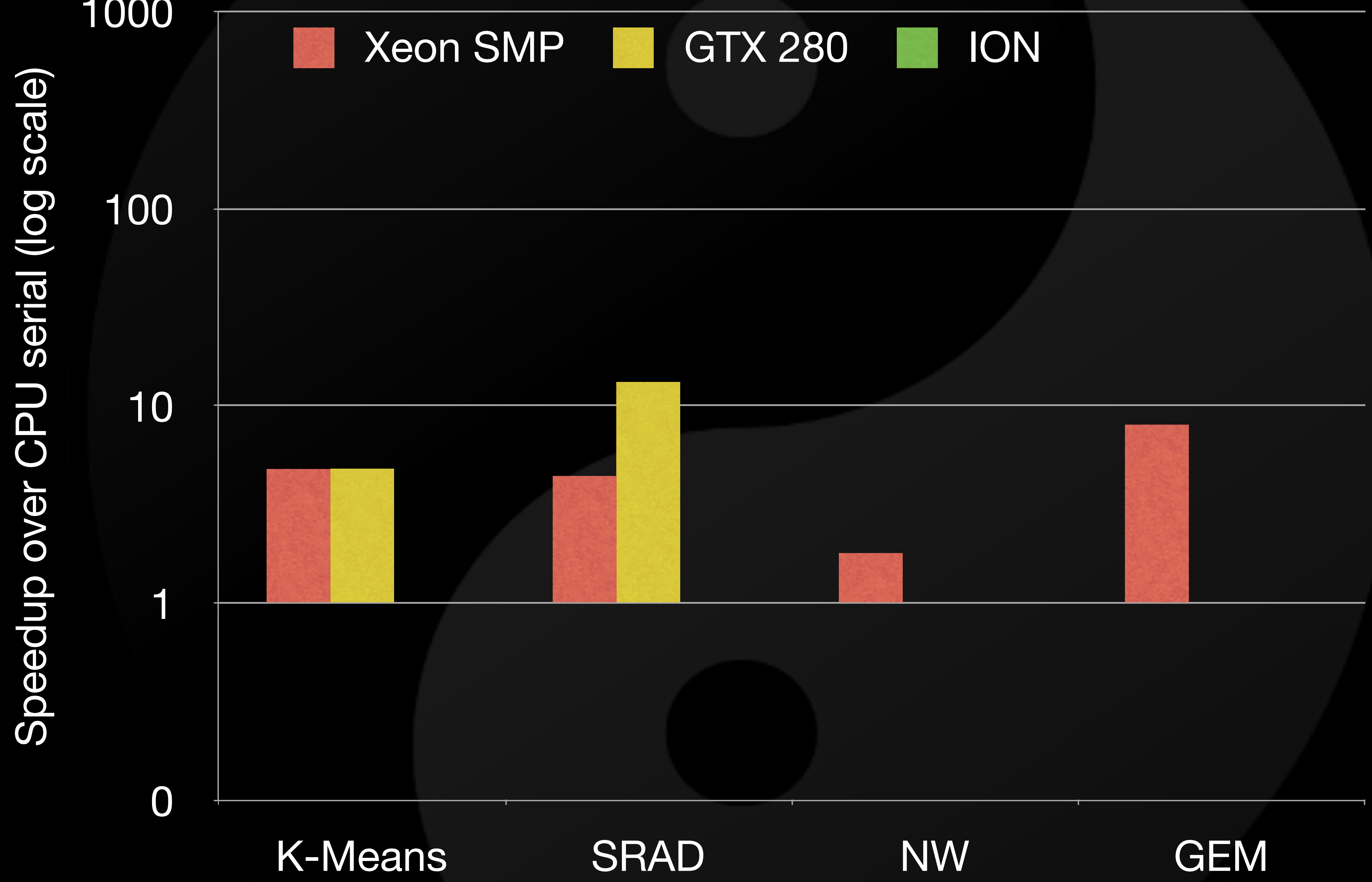
Performance



SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

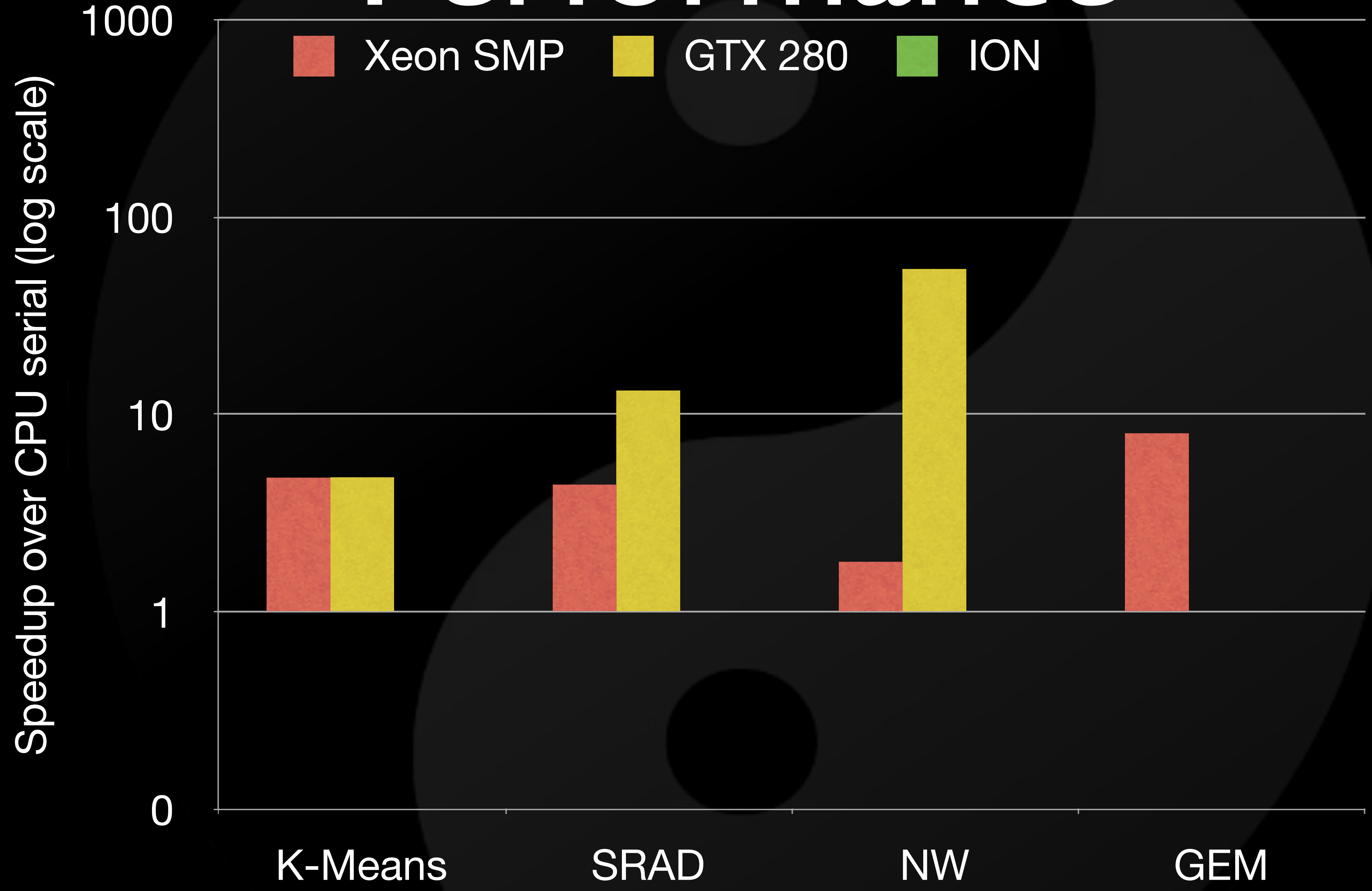
Performance



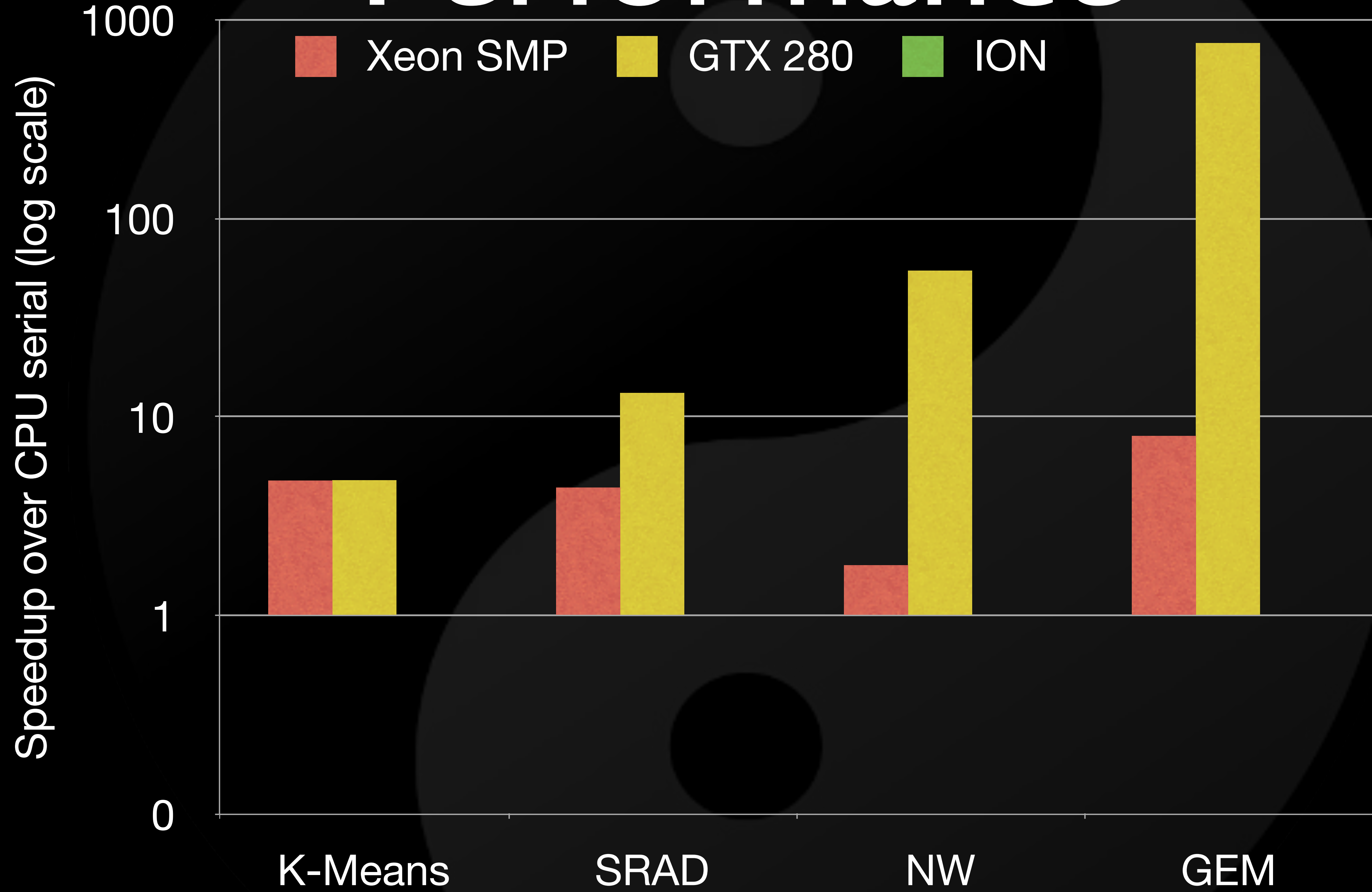
SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

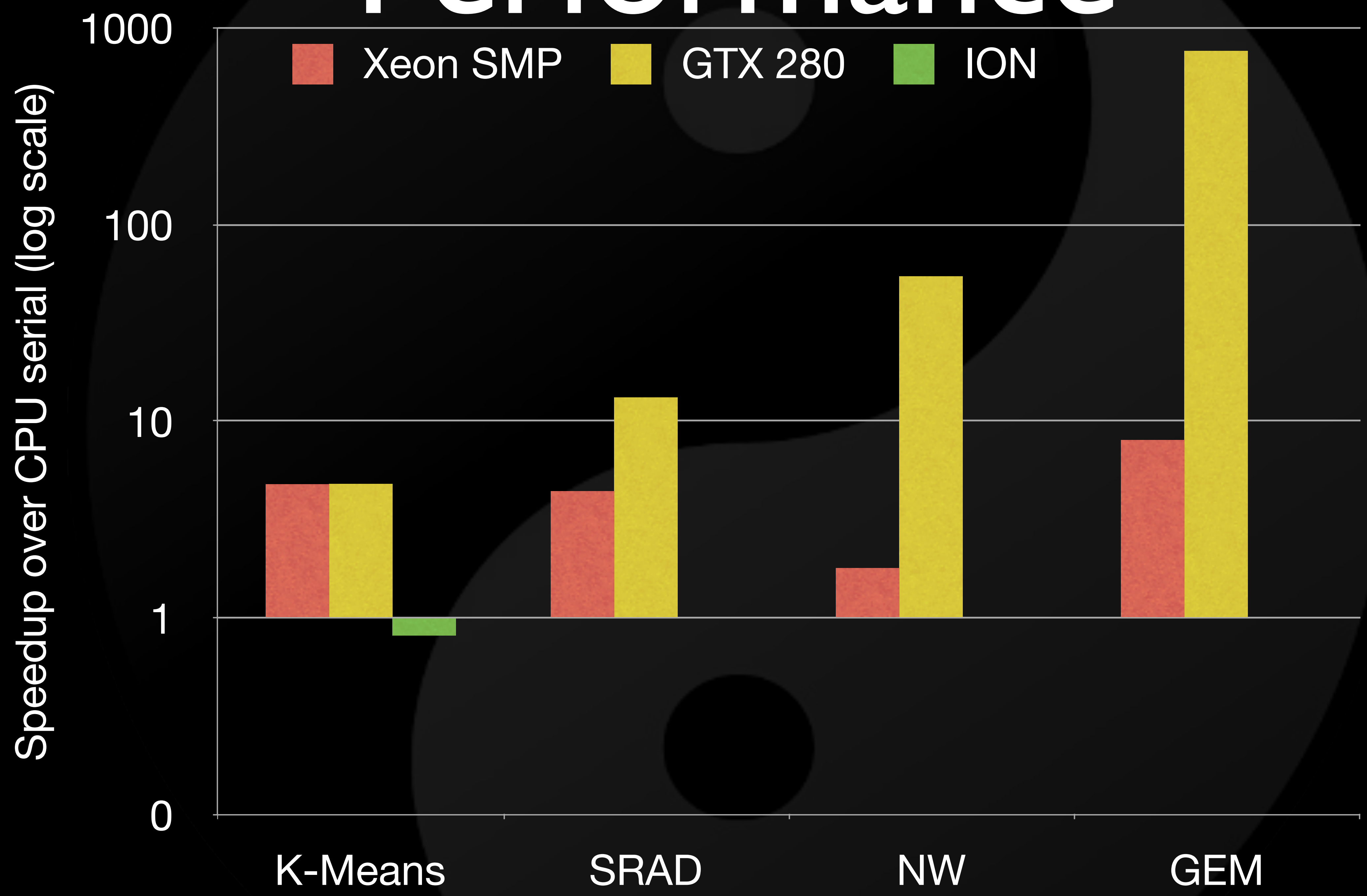
Performance



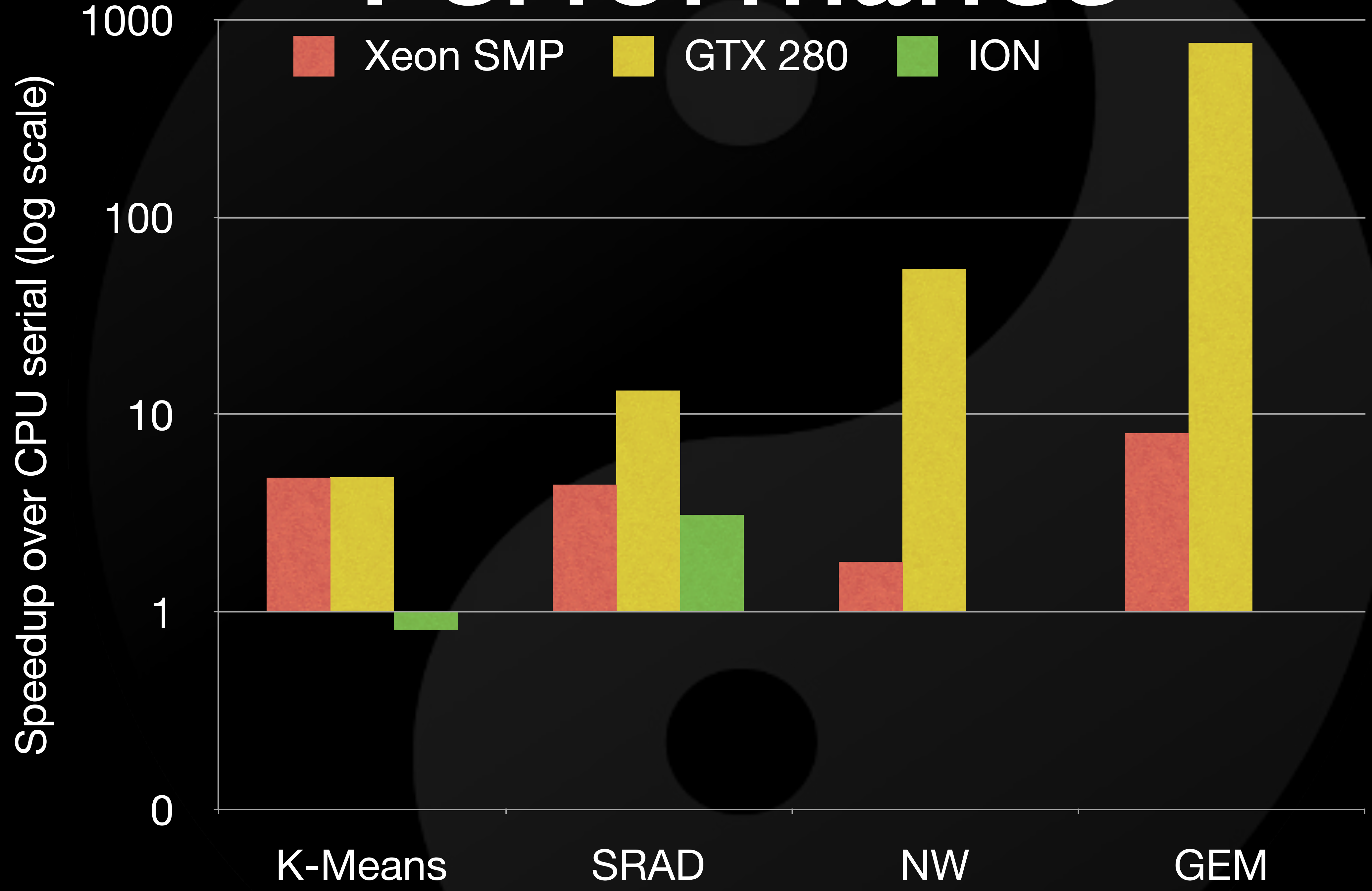
Performance



Performance



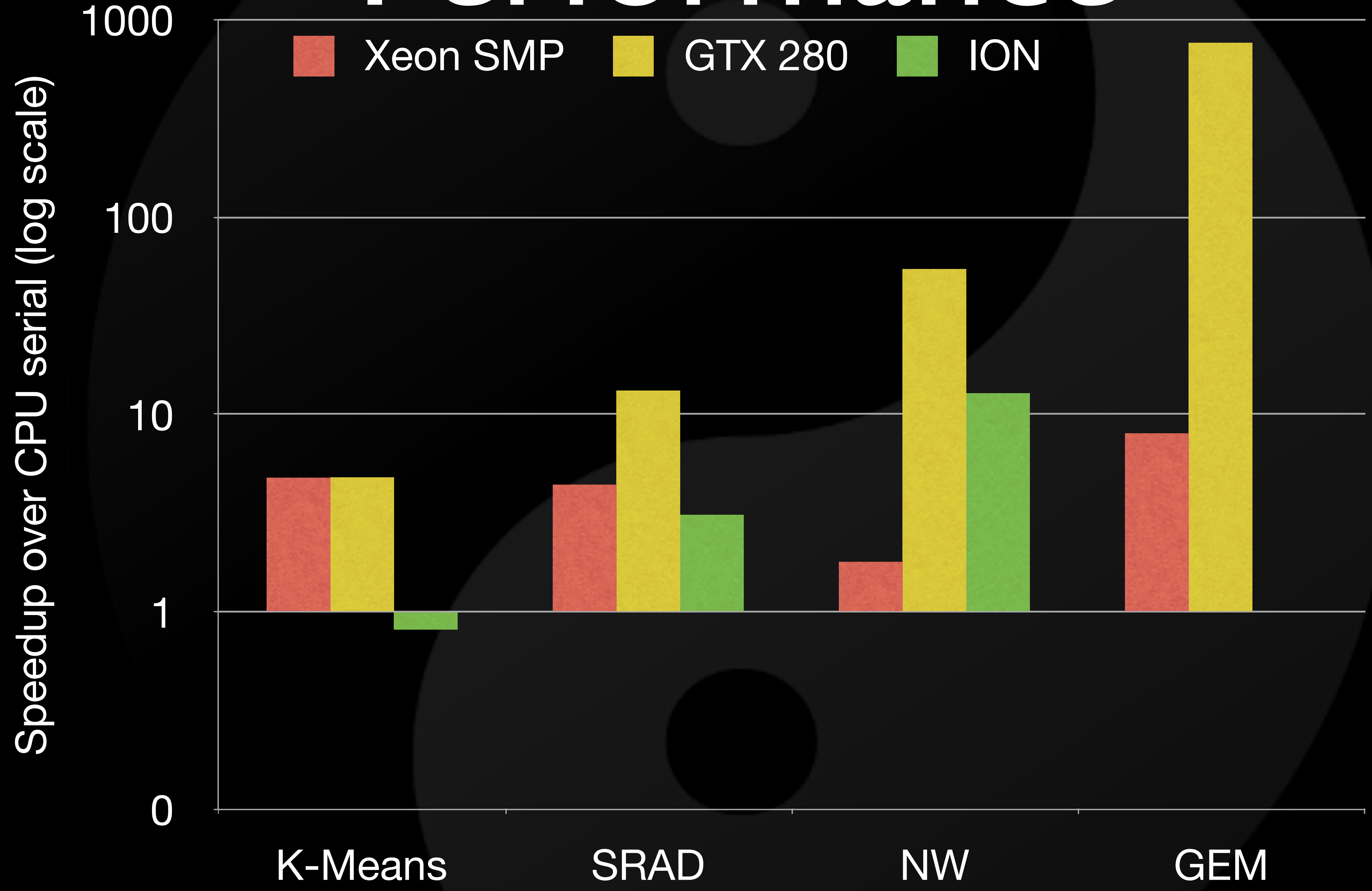
Performance



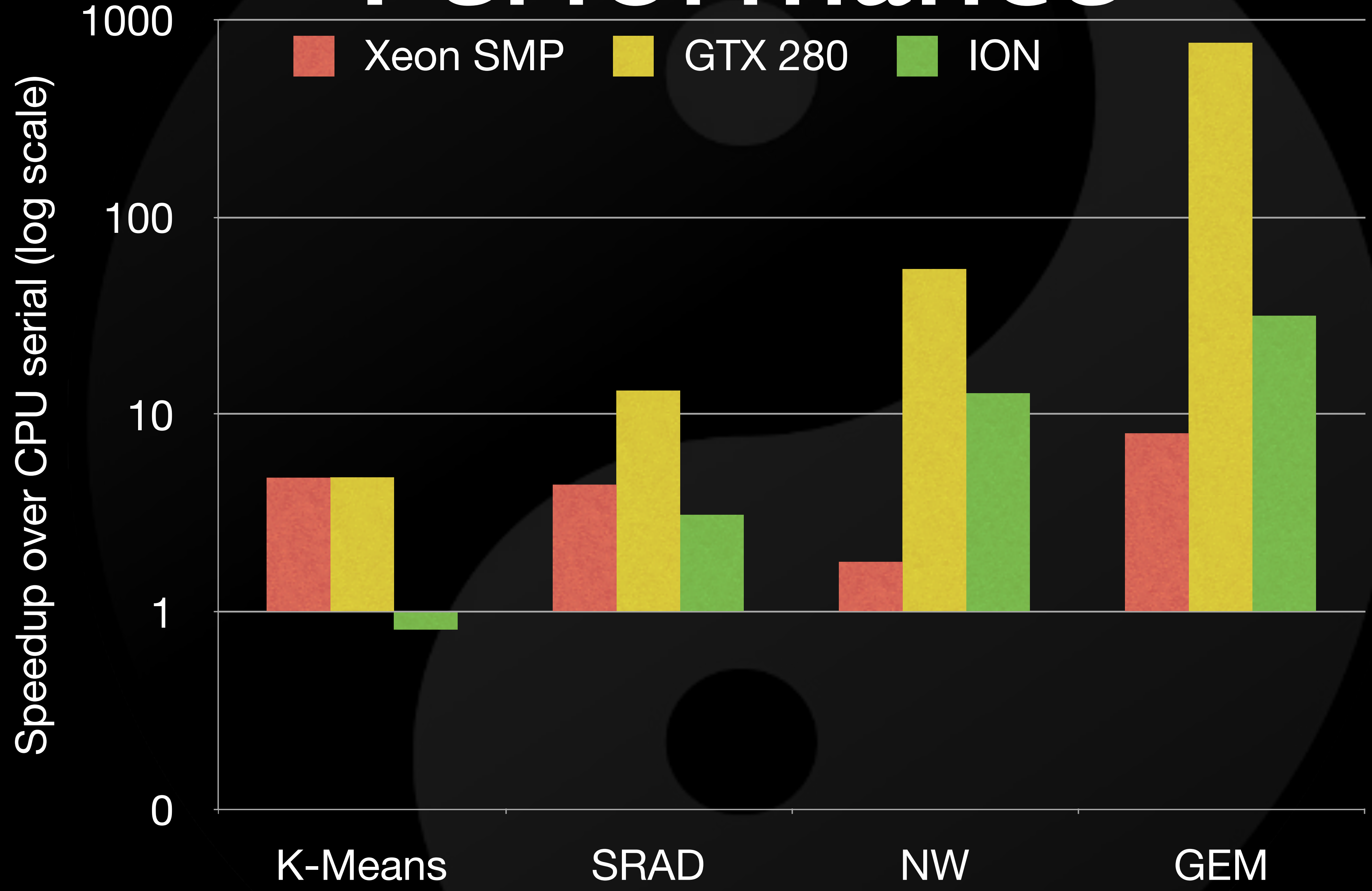
SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

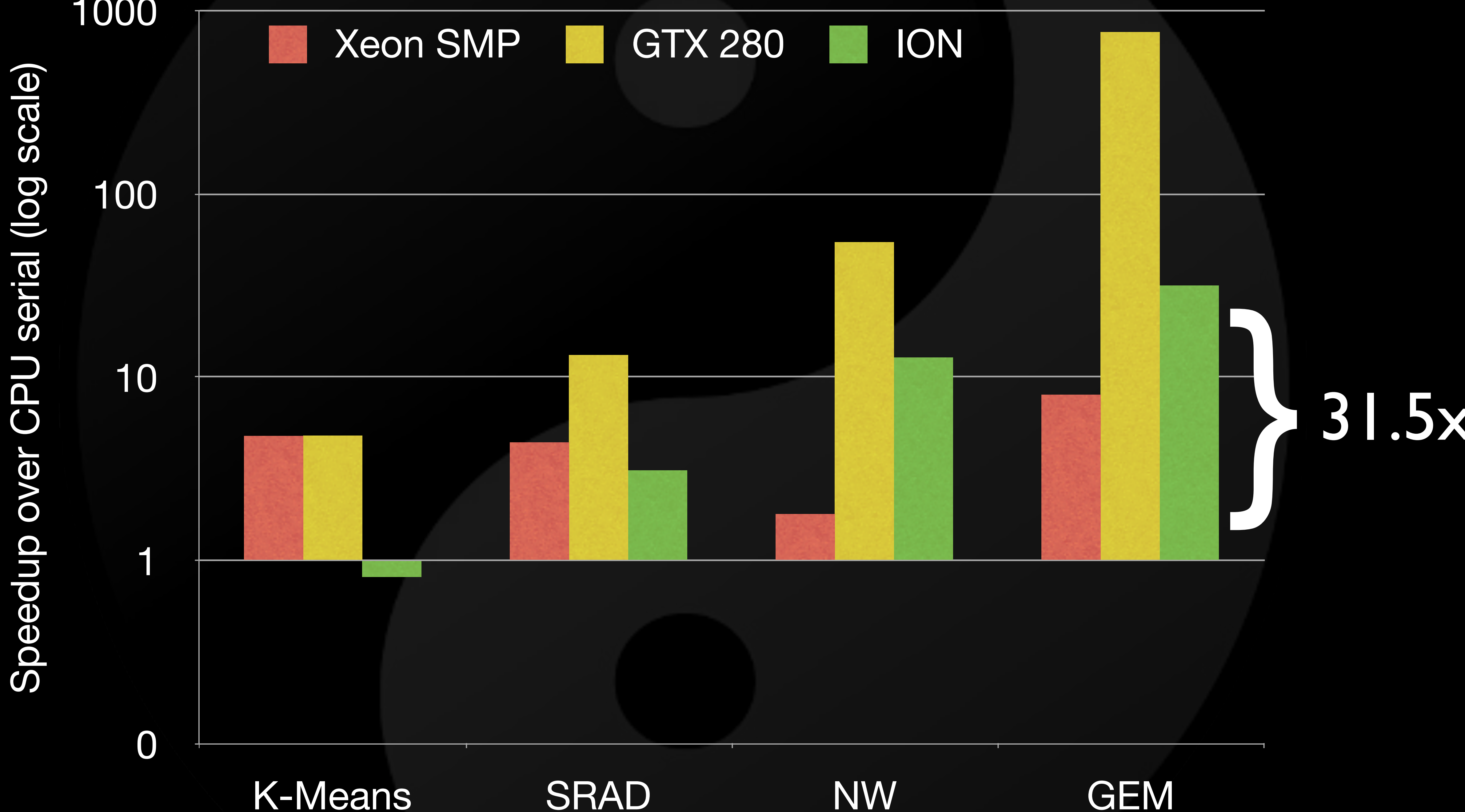
Performance



Performance

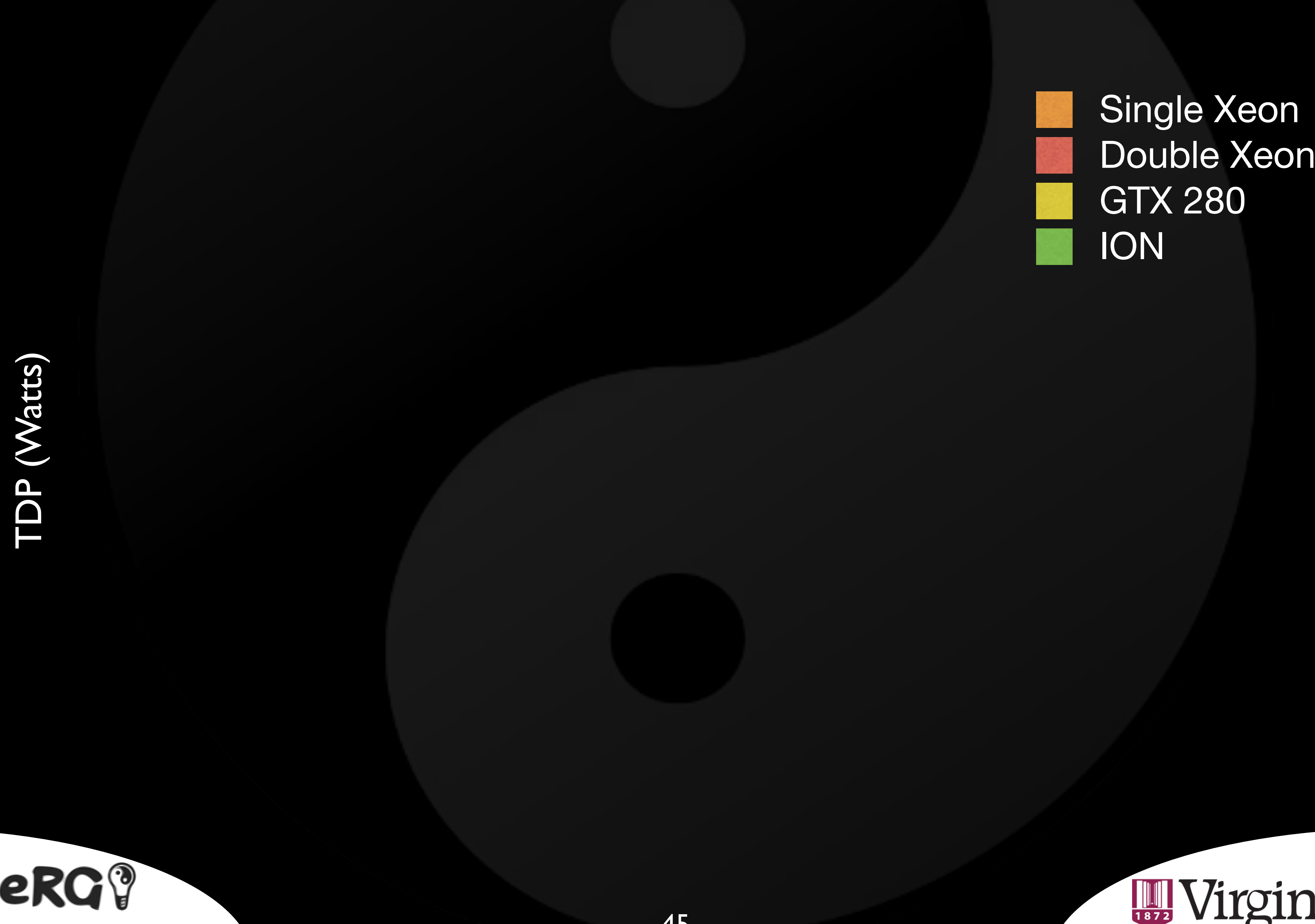


Performance

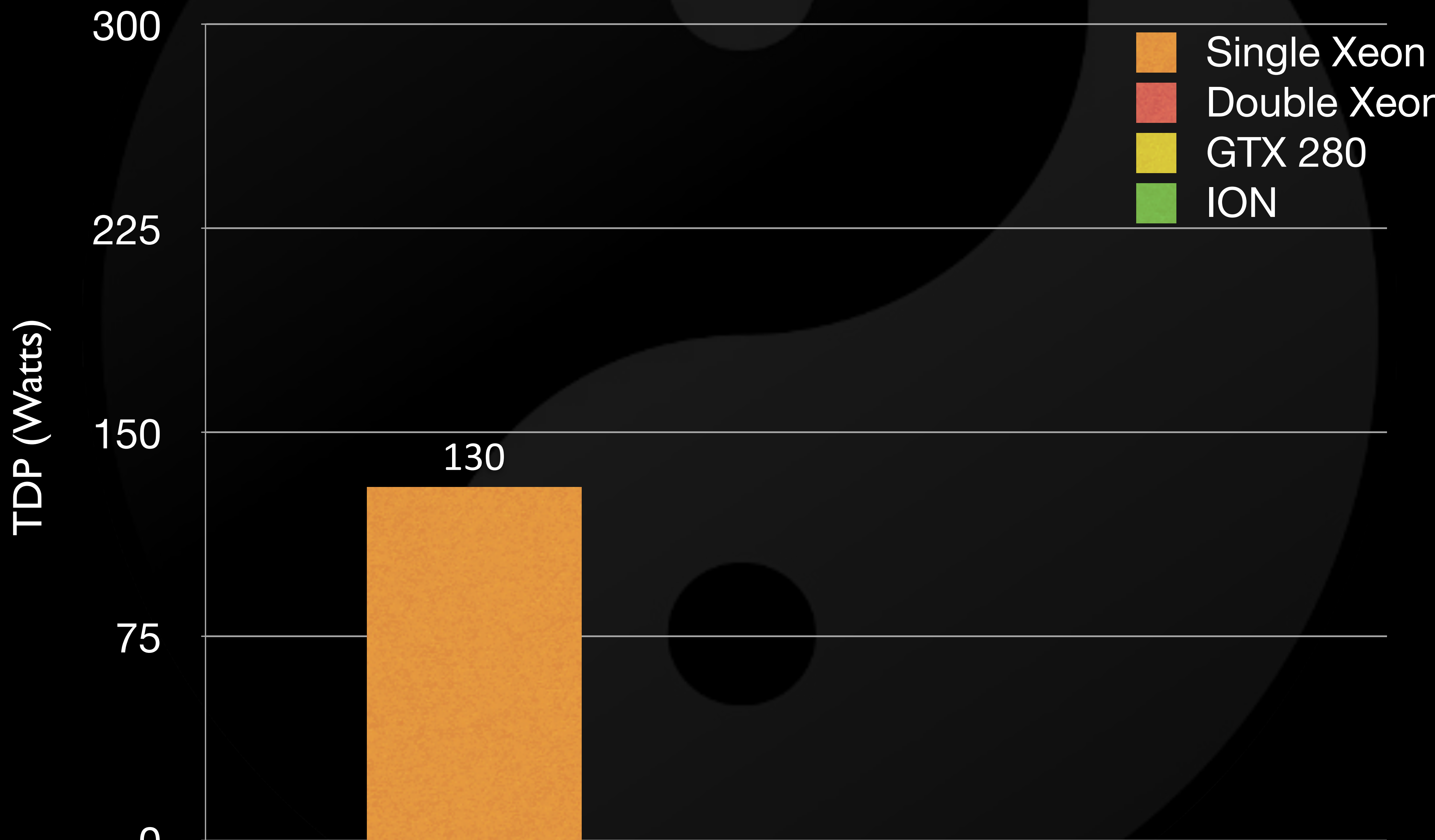


Power

Power: GPU



Power: GPU

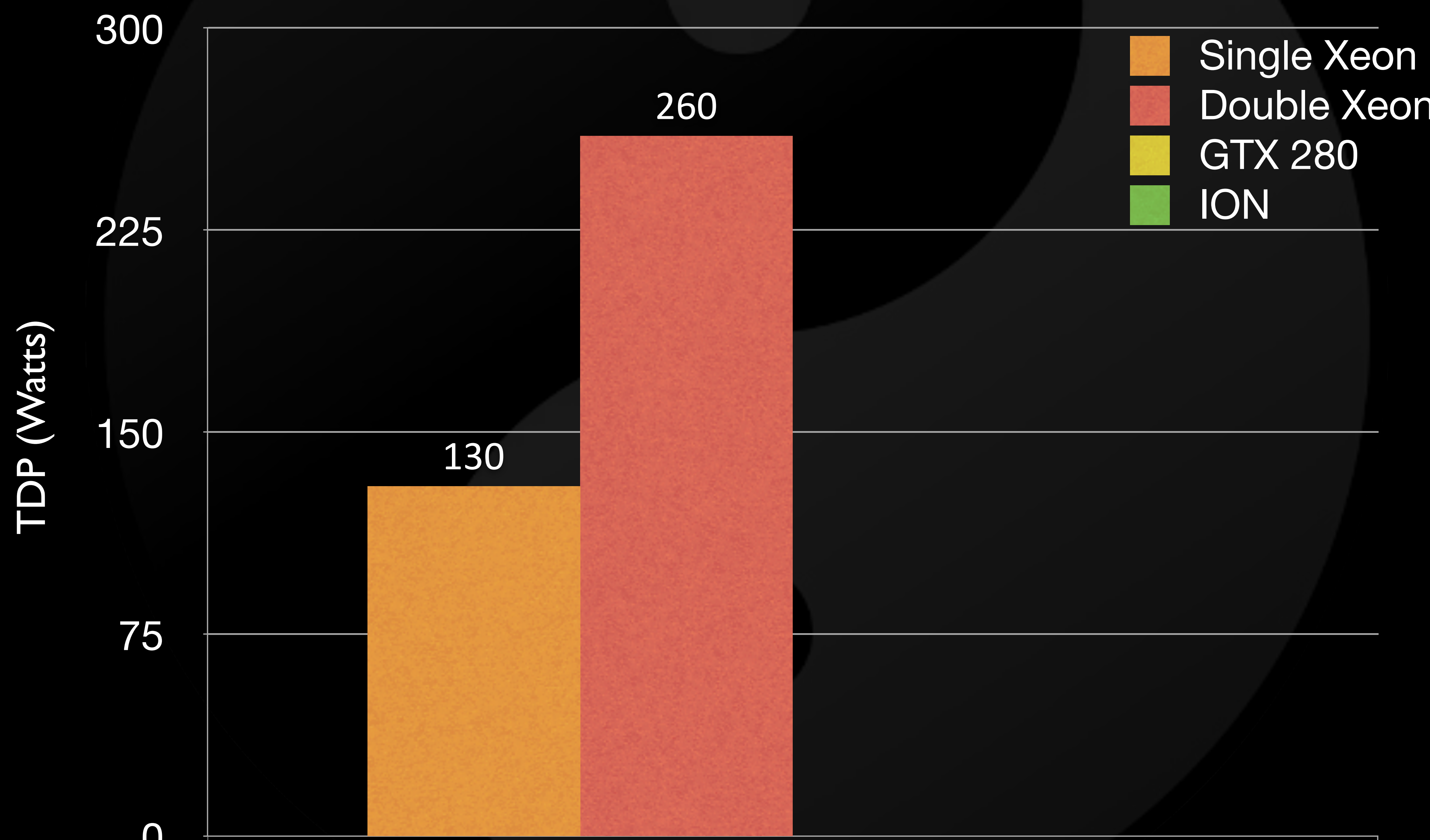


SyNeRGY
http://synergy.cs.vt.edu

45

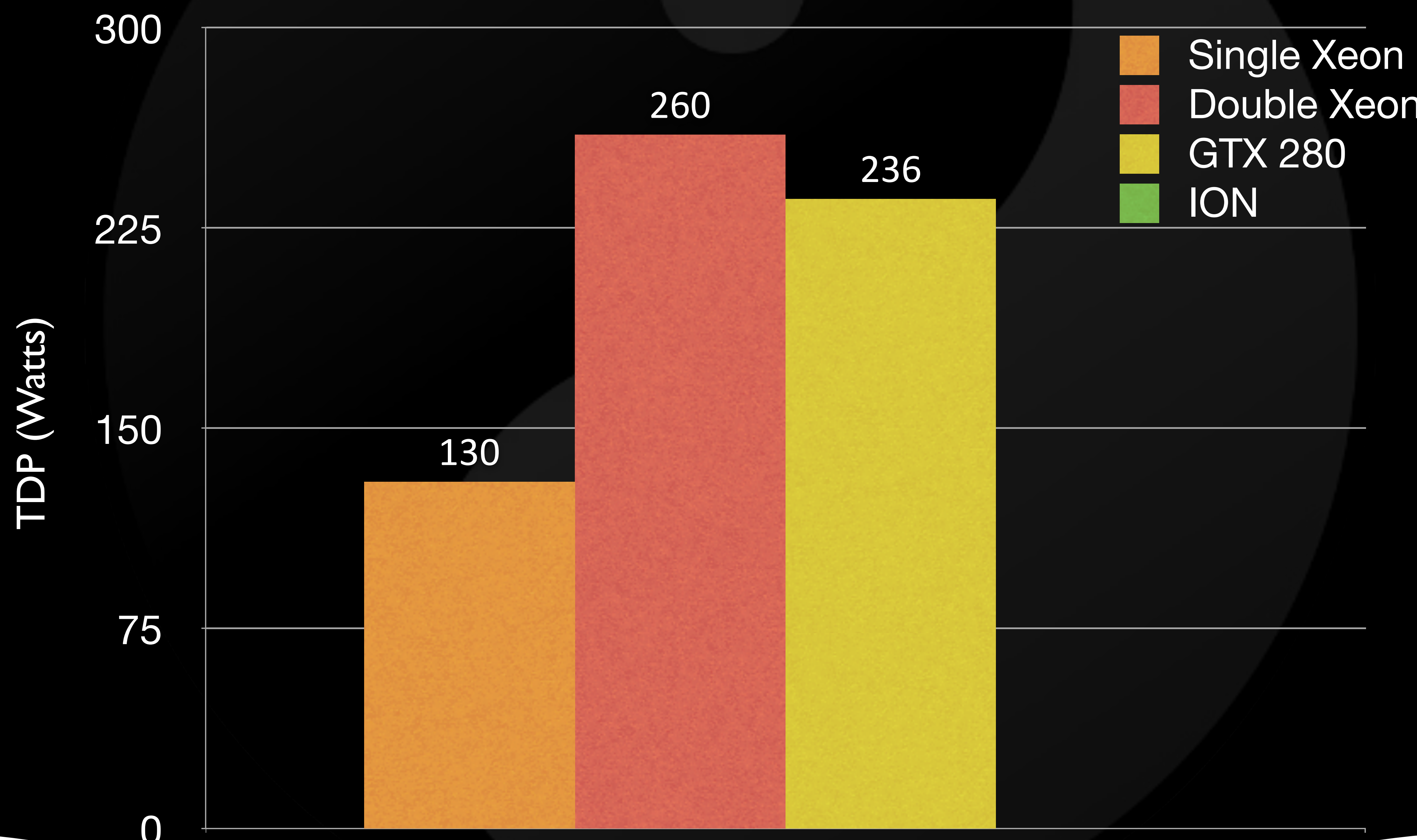
VirginiaTech
Invent the Future

Power: GPU



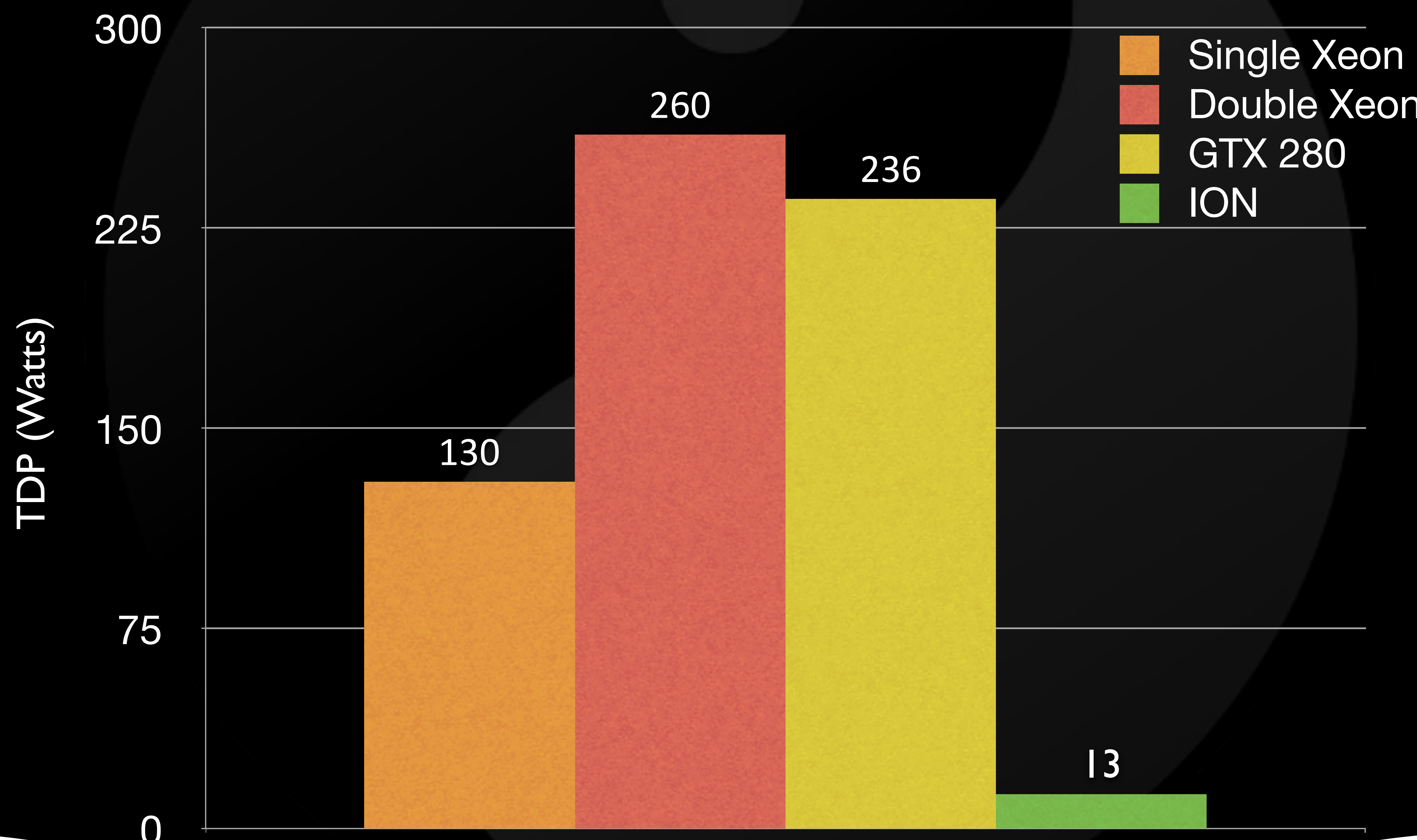
300
225
150
75
0
TDP (Watts)
260
130
Single Xeon
Double Xeon
GTX 280
ION

Power: GPU



300
225
150
75
0
TDP (Watts)
260
236
130
Single Xeon
Double Xeon
GTX 280
ION

Power: GPU

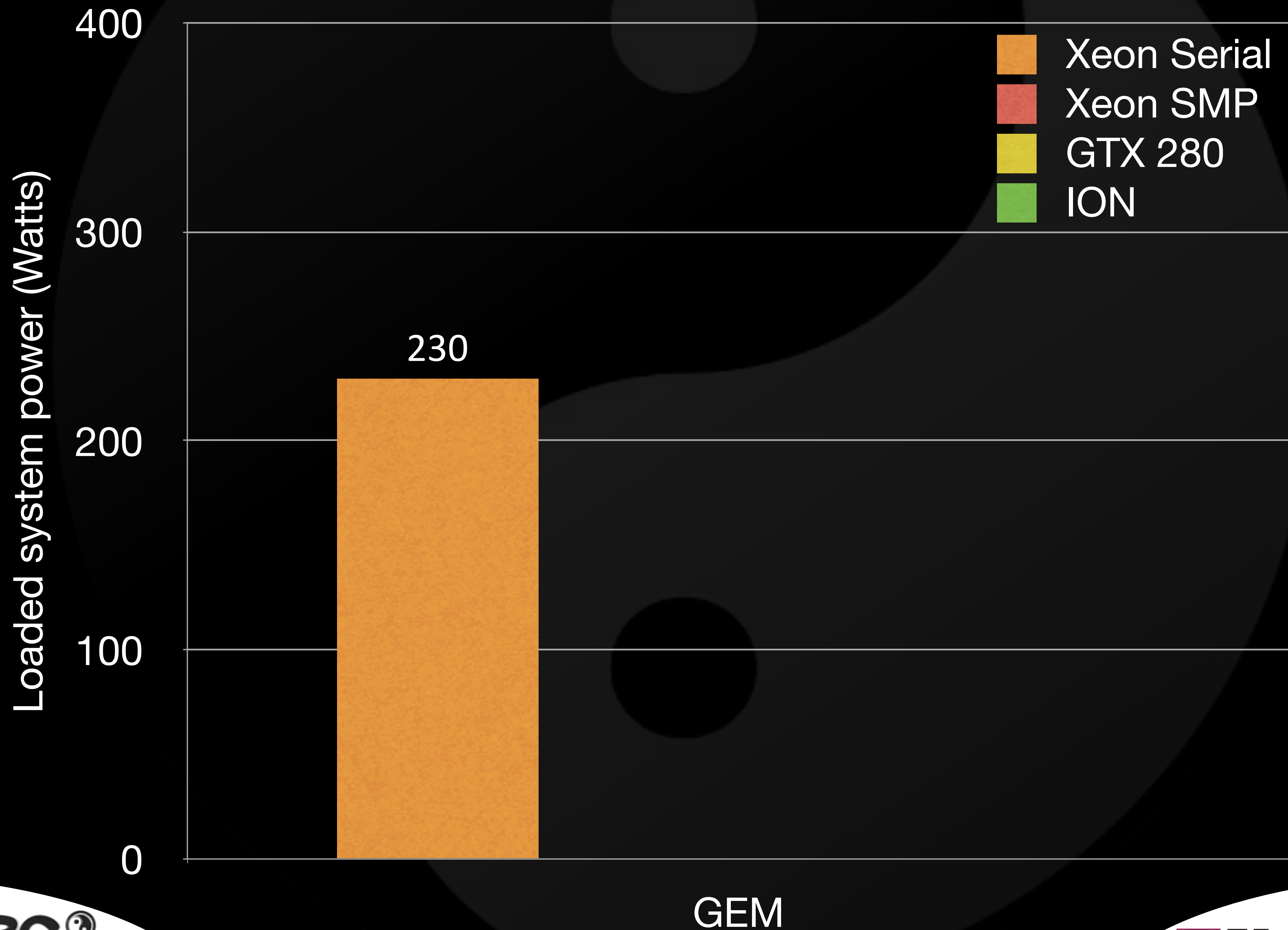


300
225
150
75
0
TDP (Watts)
130
260
236
13
Single Xeon
Double Xeon
GTX 280
ION

Power: System

Xeon Serial
Xeon SMP
GTX 280
ION

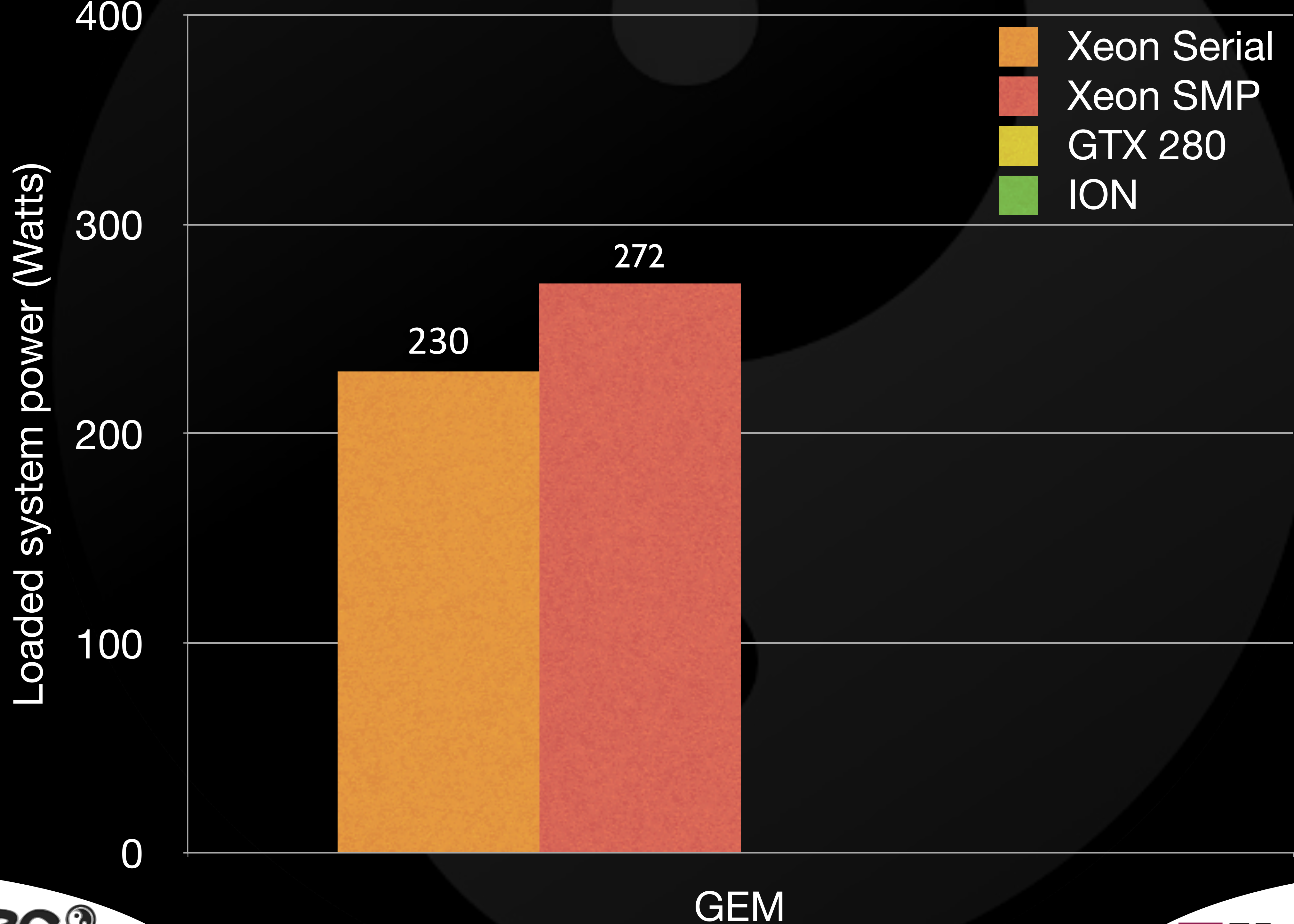
Power: System



SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
1872
Invent the Future

Power: System

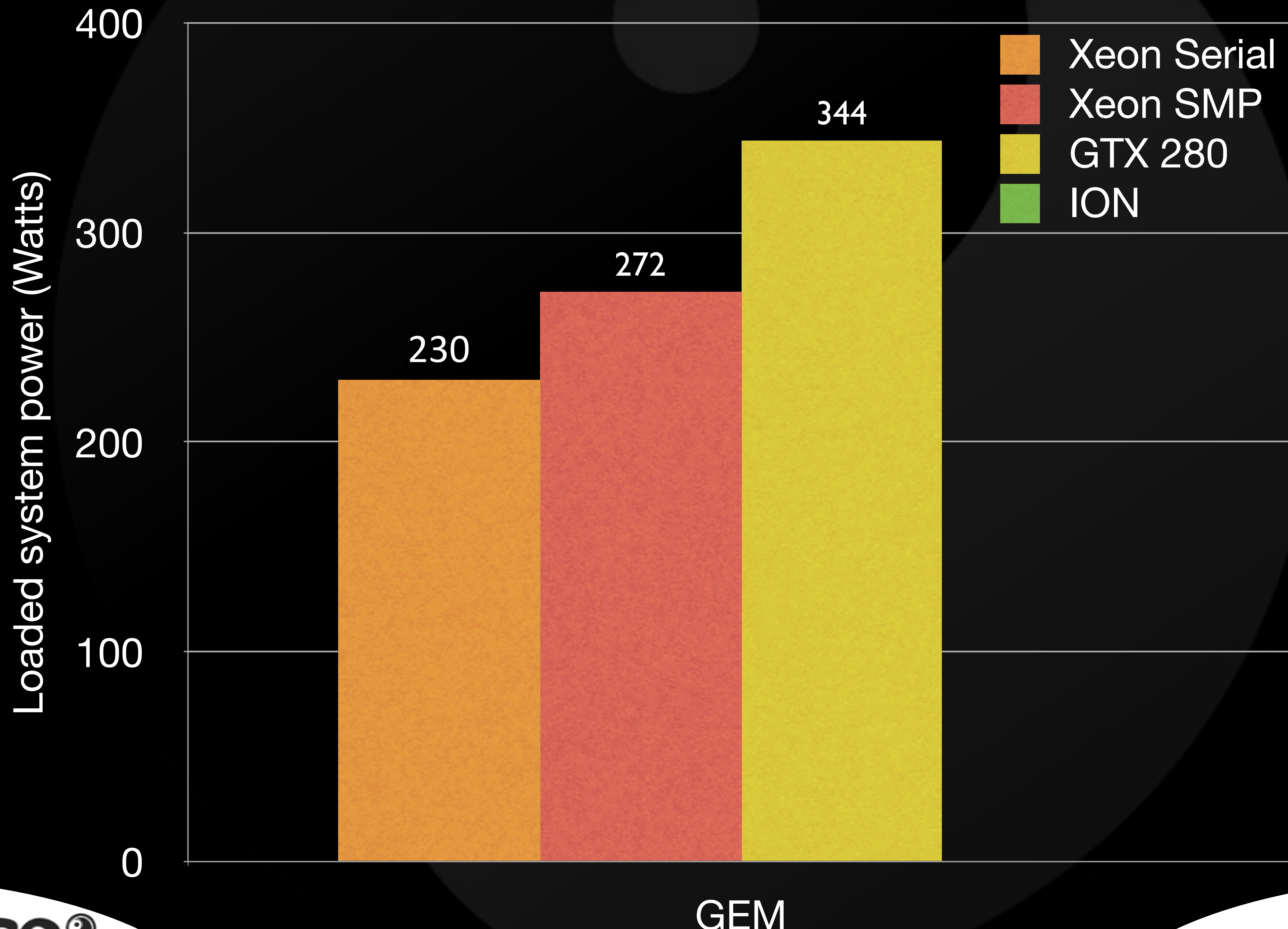


SyNeRG
http://synergy.cs.vt.edu

46

VirginiaTech
Invent the Future

Power: System



Power: System

SyNeRG
http://synergy.cs.vt.edu

46

VirginiaTech
Invent the Future

Energy Efficiency

Energy Efficiency

Energy Efficiency

- Metric: Energy Delay Product (EDP)
  - Common circuit design metric

Energy Efficiency

- Metric: Energy Delay Product (EDP)
  - Common circuit design metric
- EDP = (joules) * (seconds)

Energy Efficiency

- Metric: Energy Delay Product (EDP)
  - Common circuit design metric
- EDP = (joules) * (seconds)
- EDP = (Watts * seconds) * (seconds)

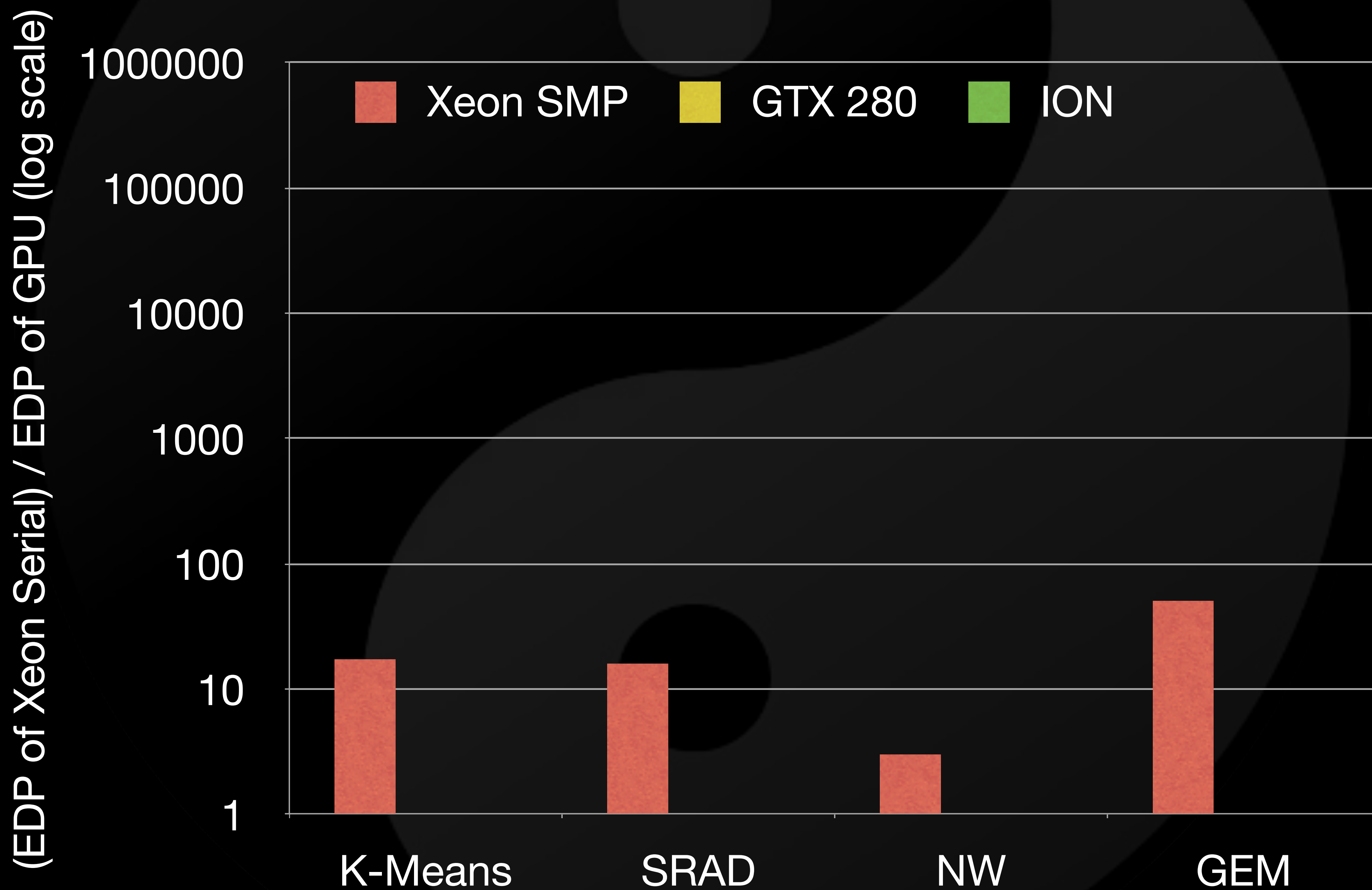
Energy Efficiency

- Metric: Energy Delay Product (EDP)
  - Common circuit design metric
- EDP = (joules) * (seconds)
- EDP = (Watts * seconds) * (seconds)
- Performance-centric energy efficiency

Energy Efficiency

Xeon SMP GTX 280 ION

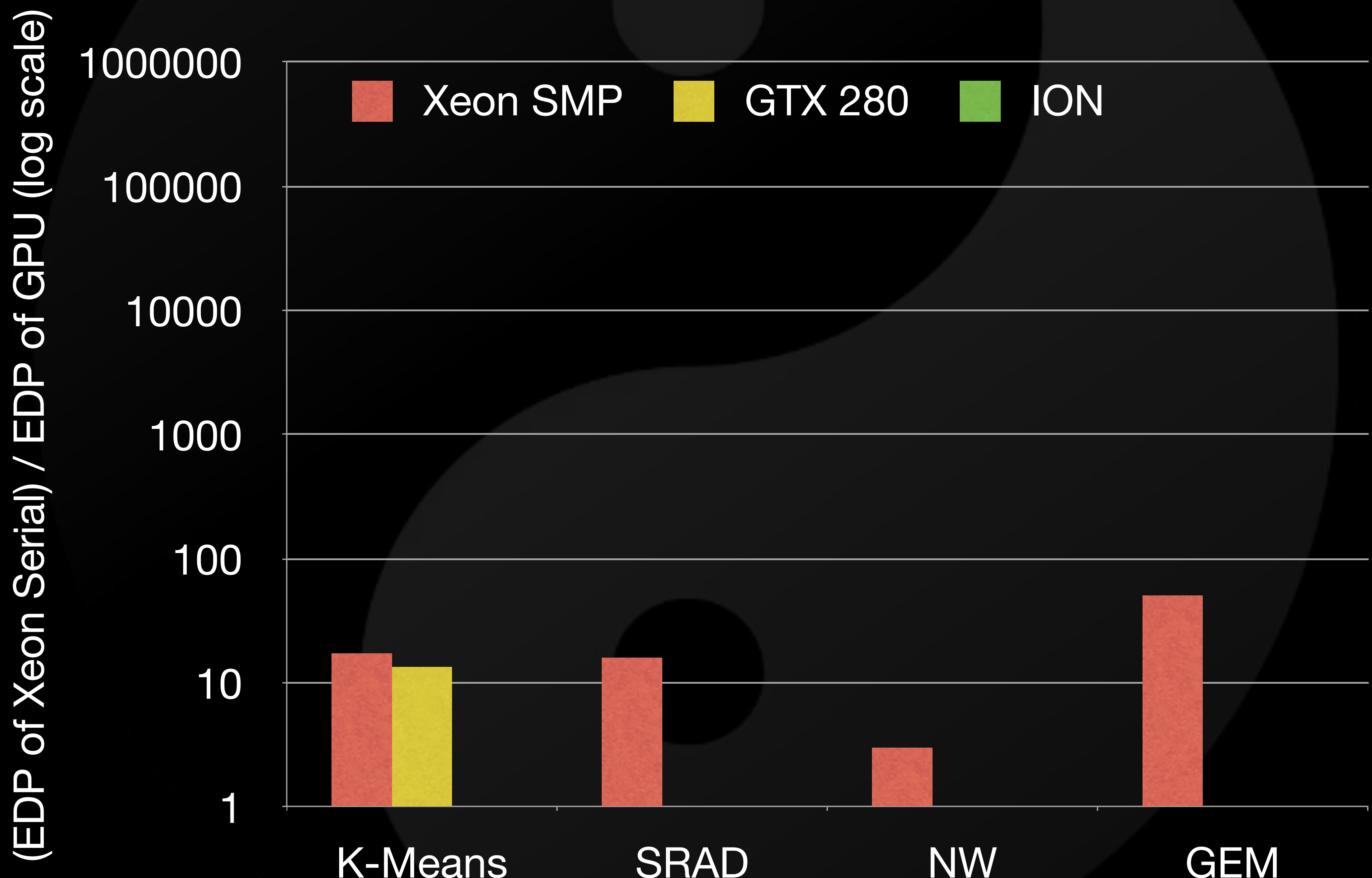
Energy Efficiency



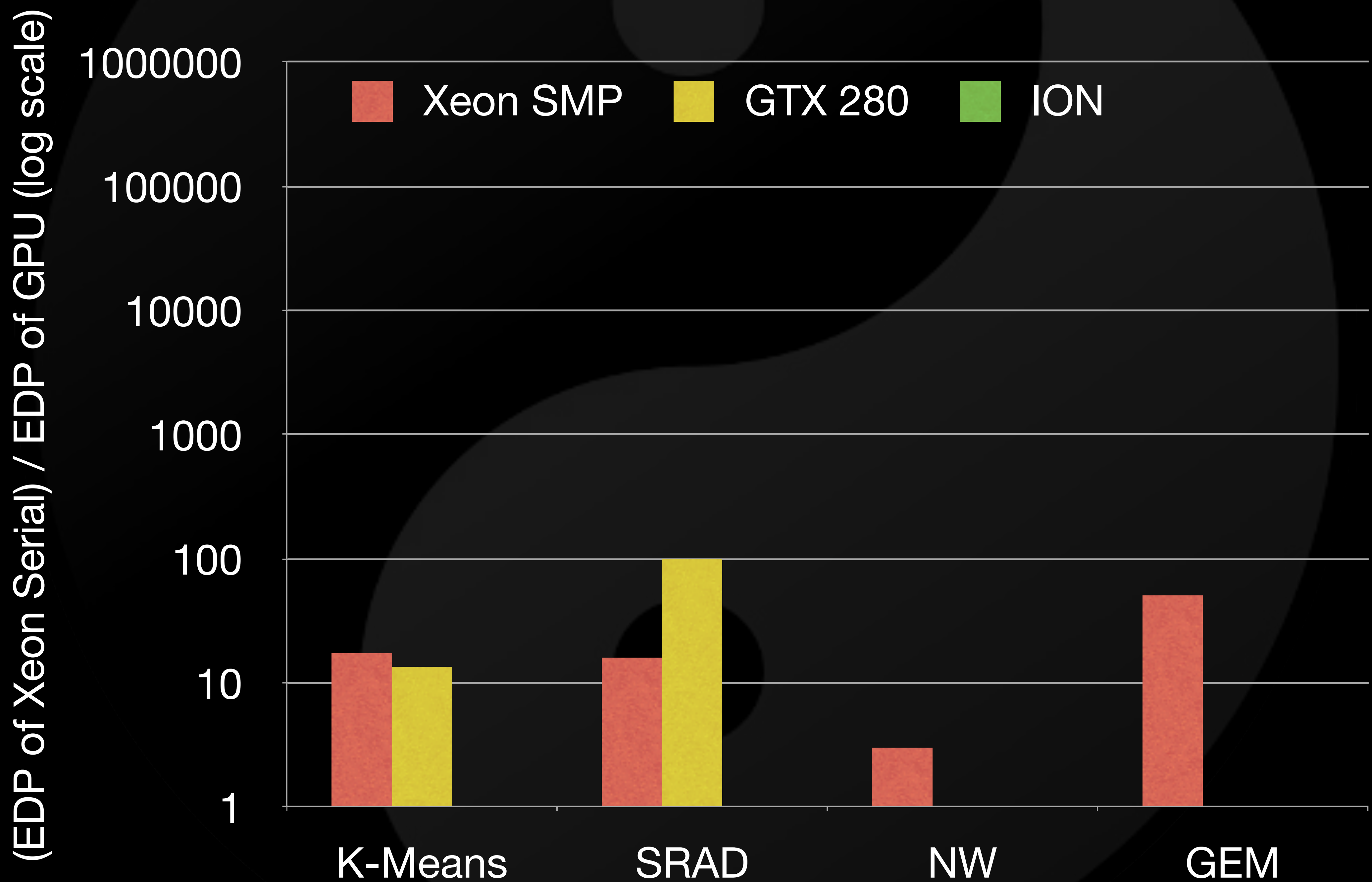
SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

Energy Efficiency



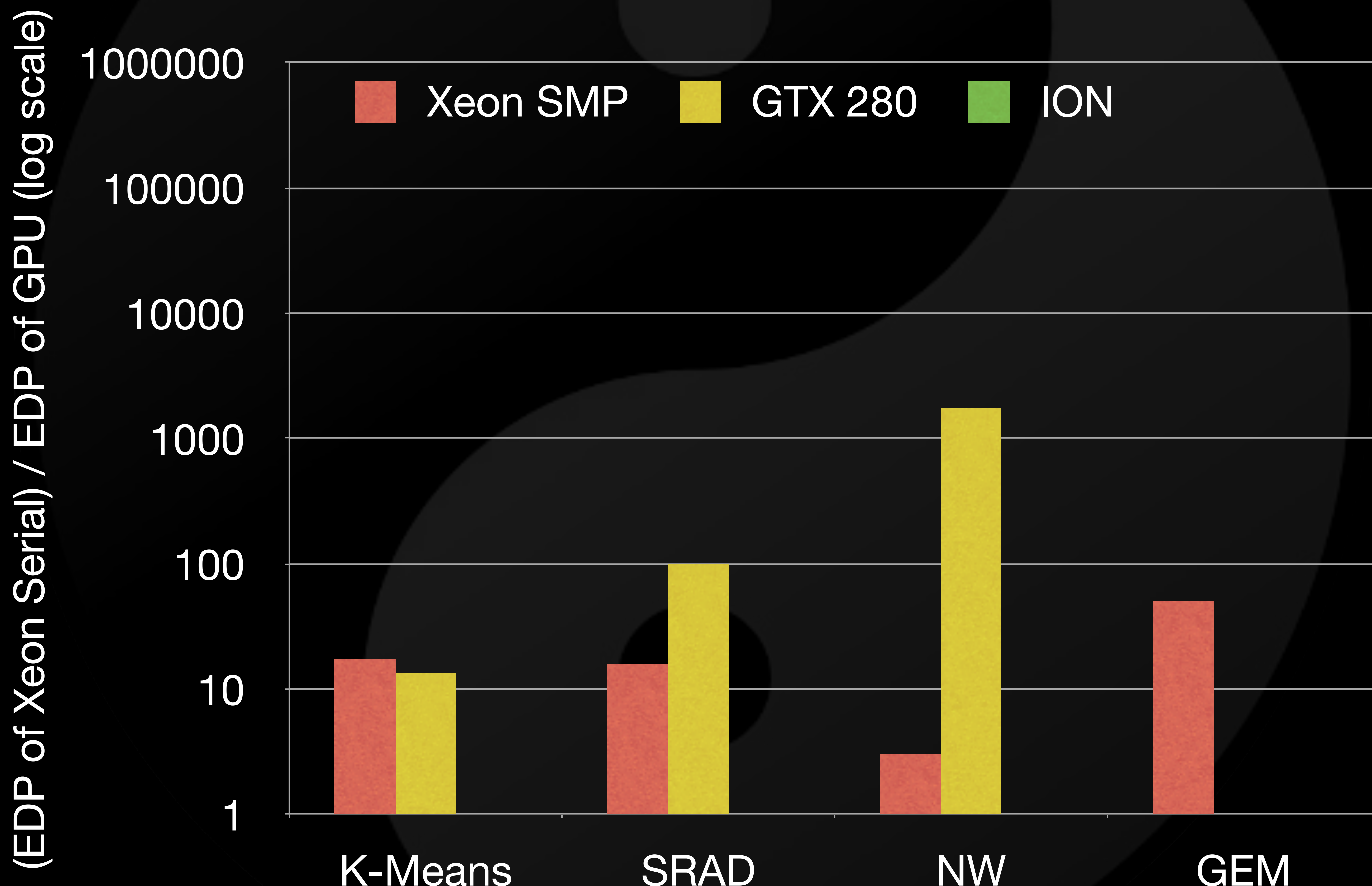
Energy Efficiency



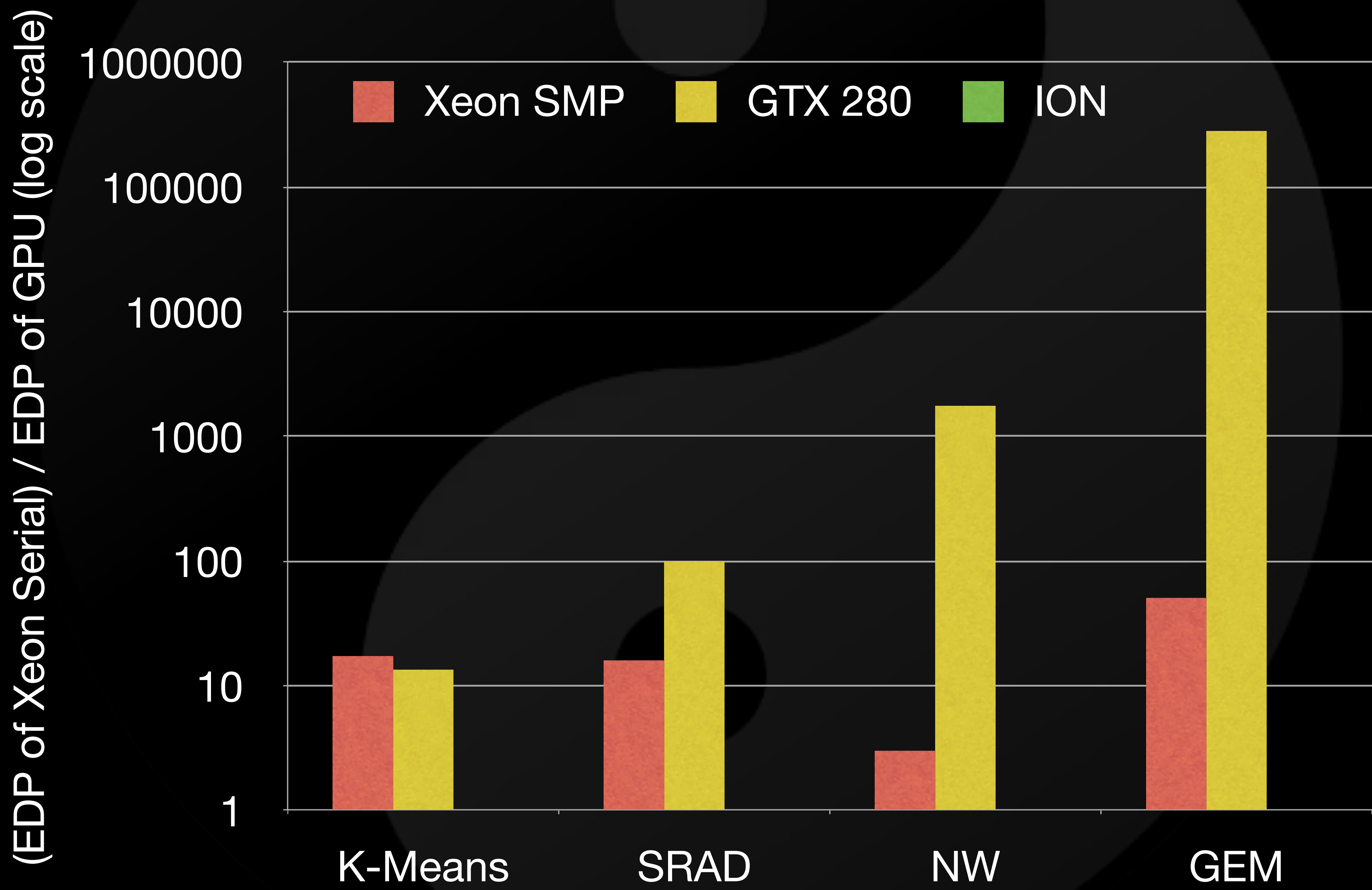
SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

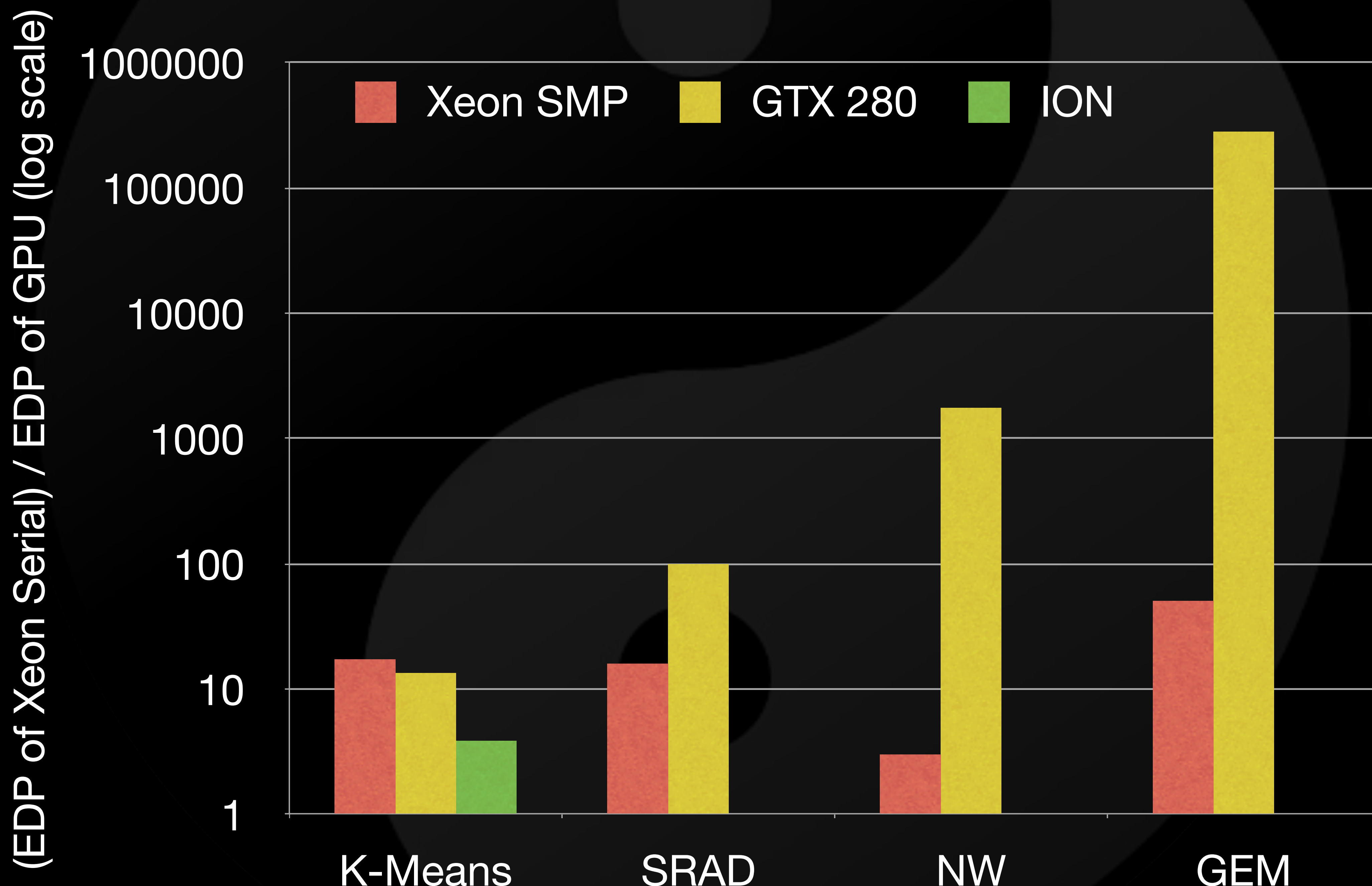
Energy Efficiency



Energy Efficiency



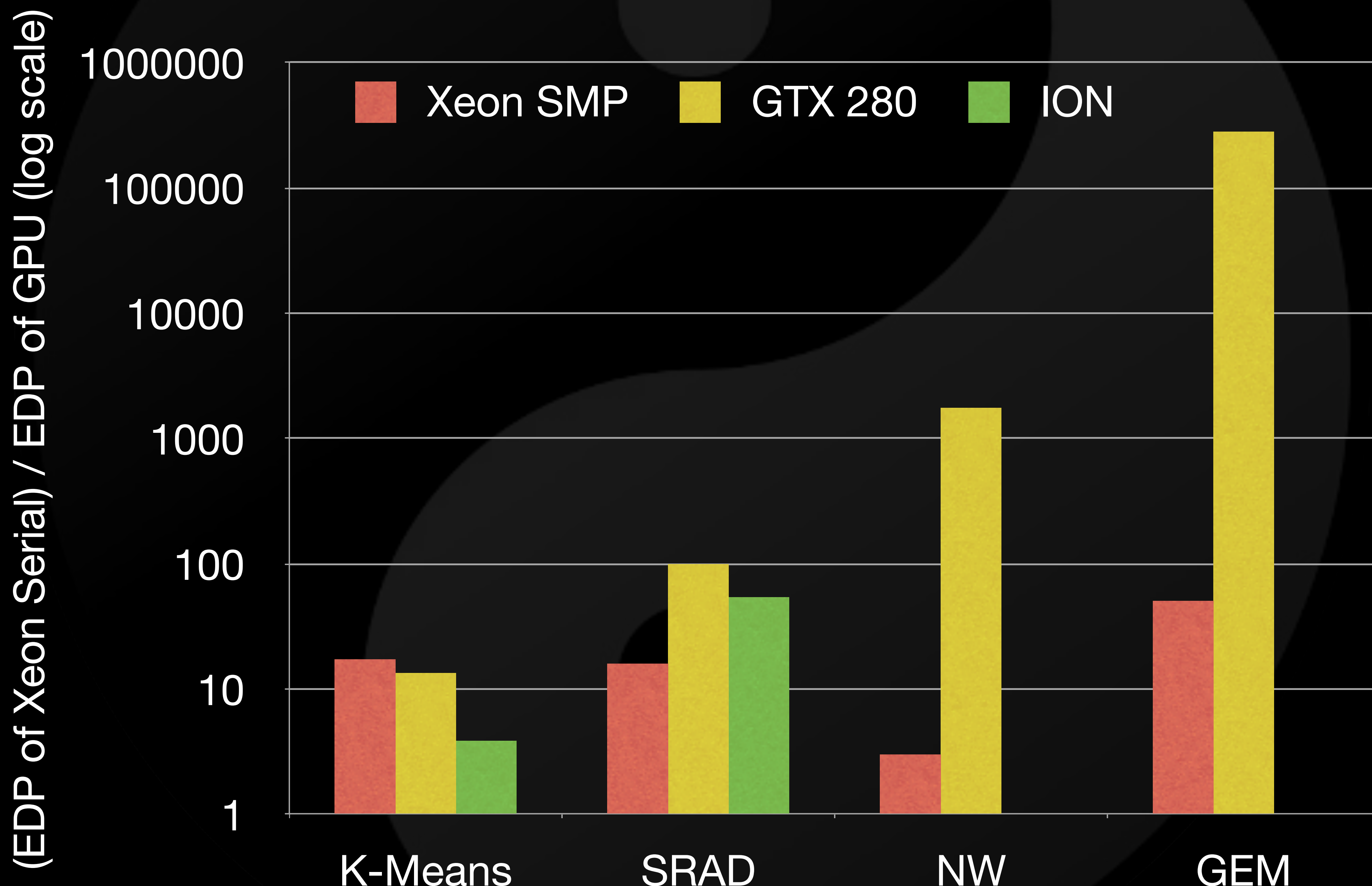
Energy Efficiency



SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

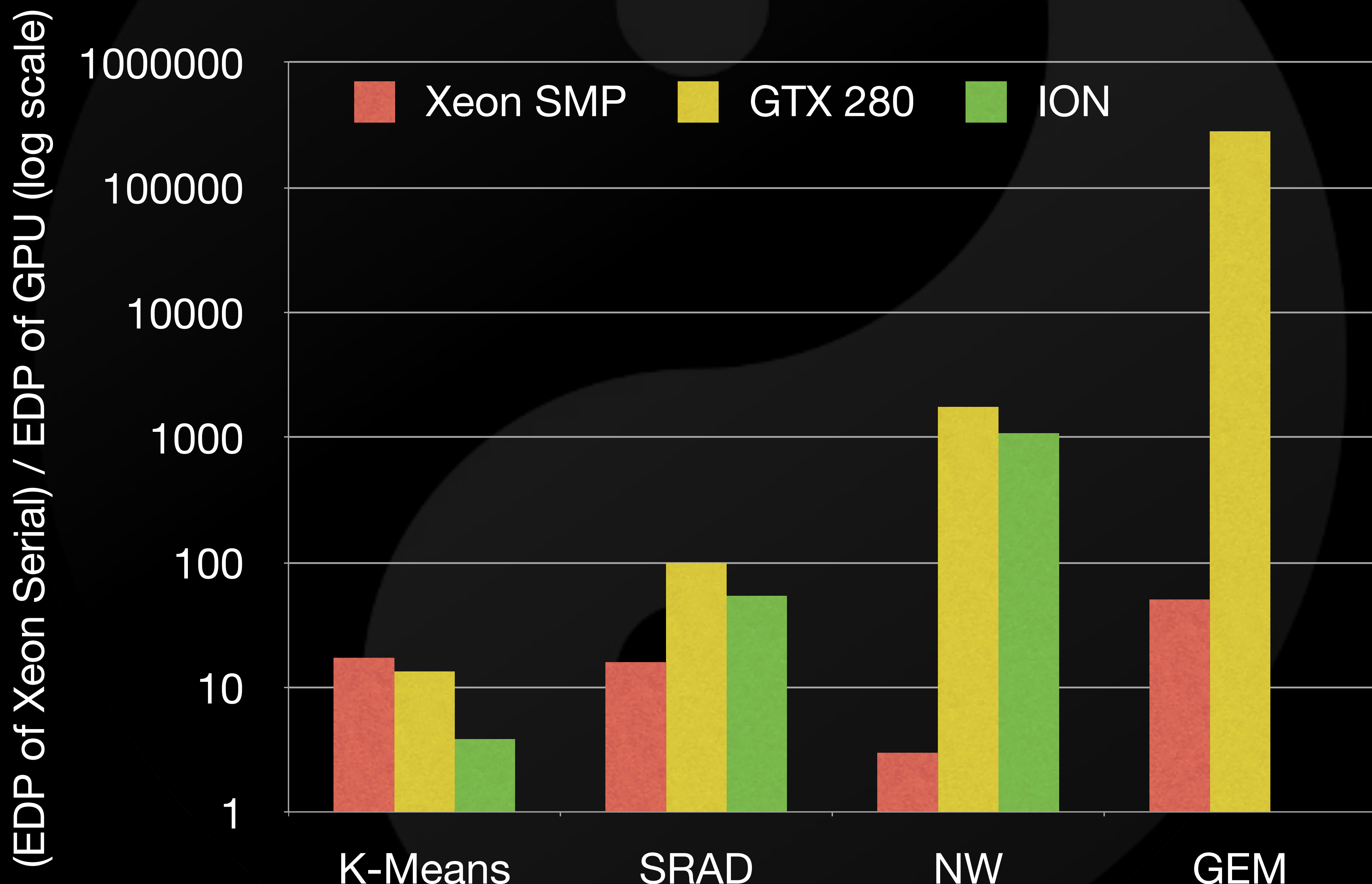
Energy Efficiency



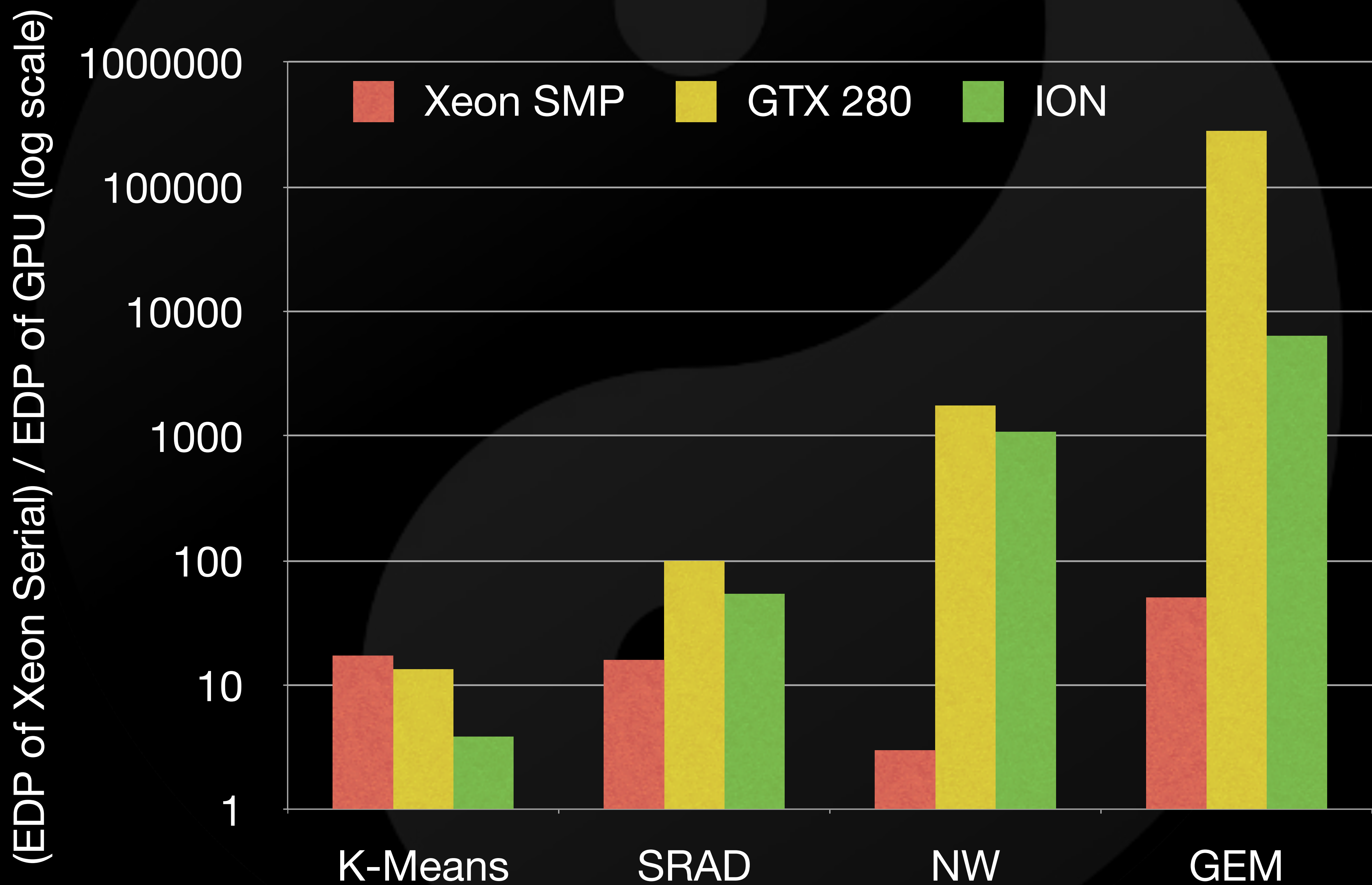
SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

Energy Efficiency



Energy Efficiency



SyNeRG
http://synergy.cs.vt.edu

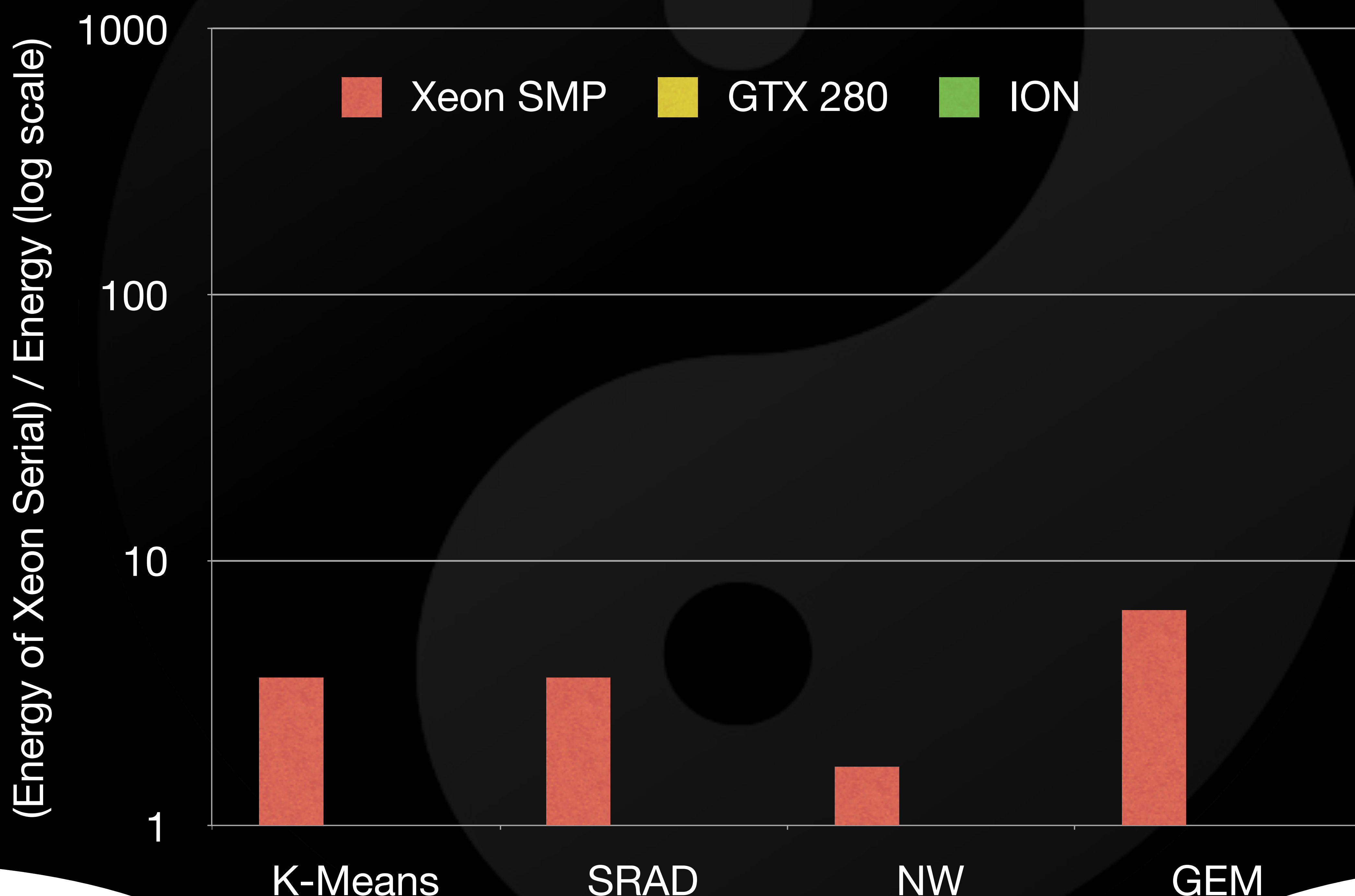
49

VirginiaTech
Invent the Future

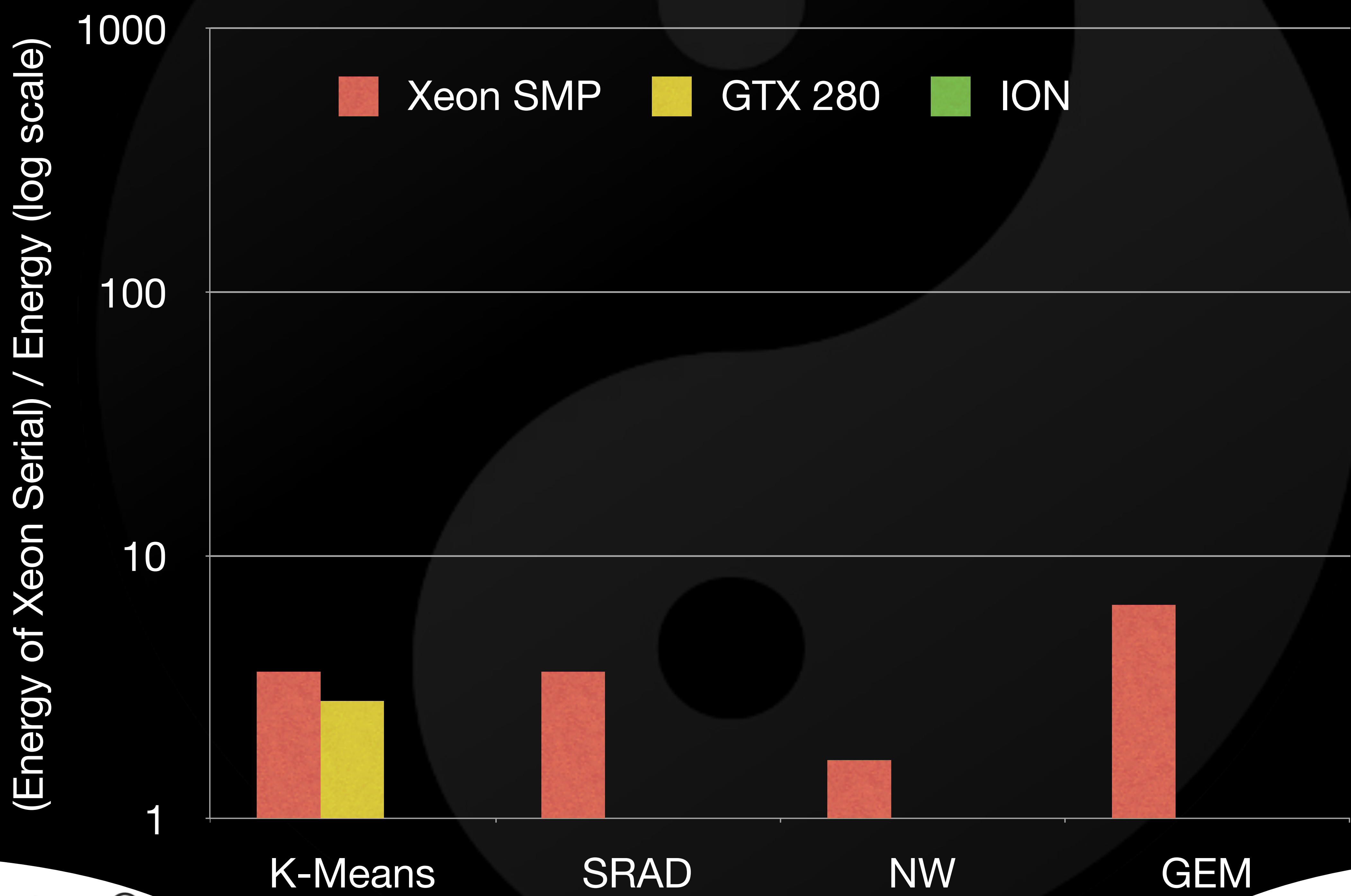
Energy Overall

Xeon SMP GTX 280 ION

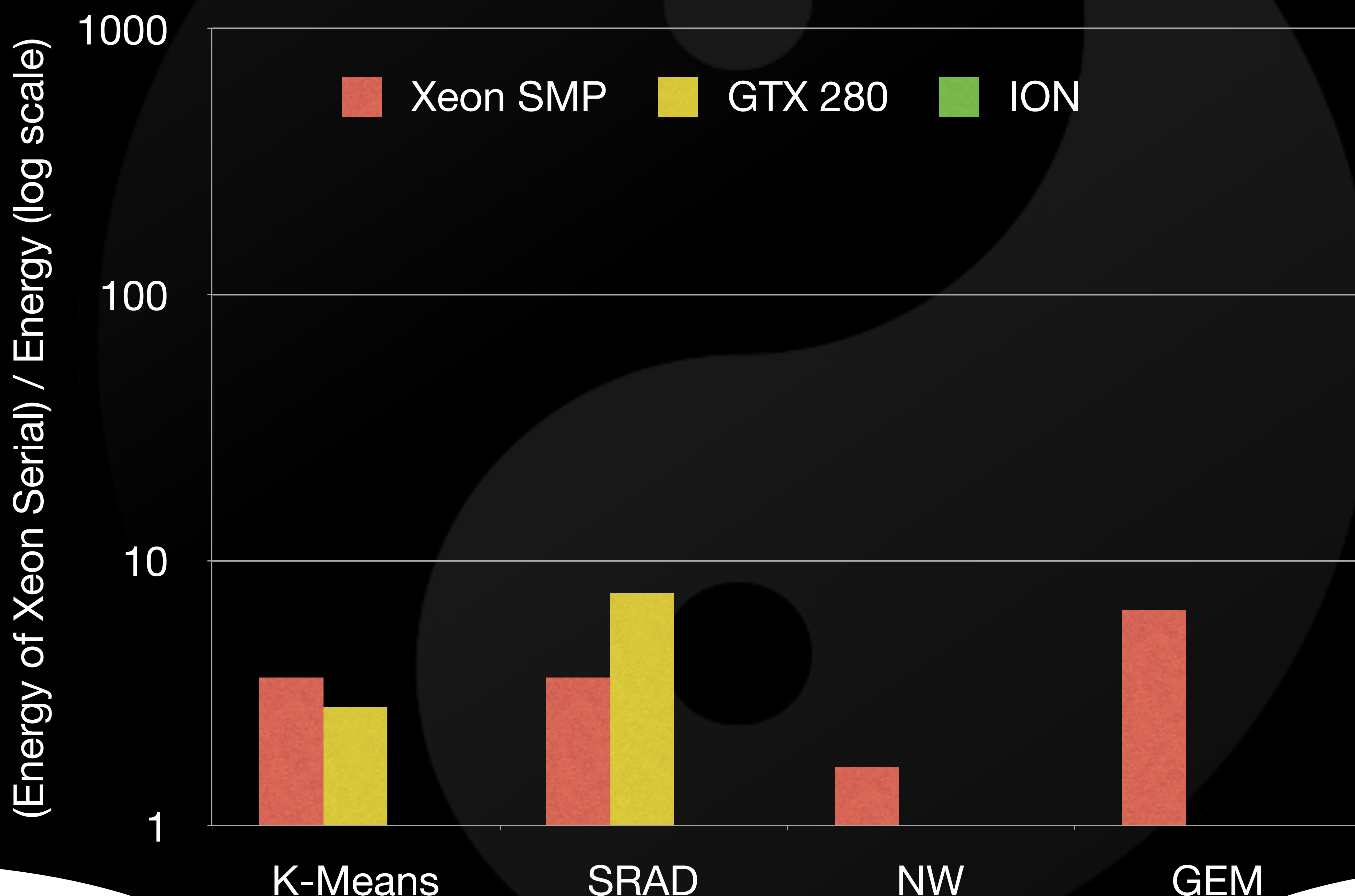
Energy Overall



Energy Overall



Energy Overall

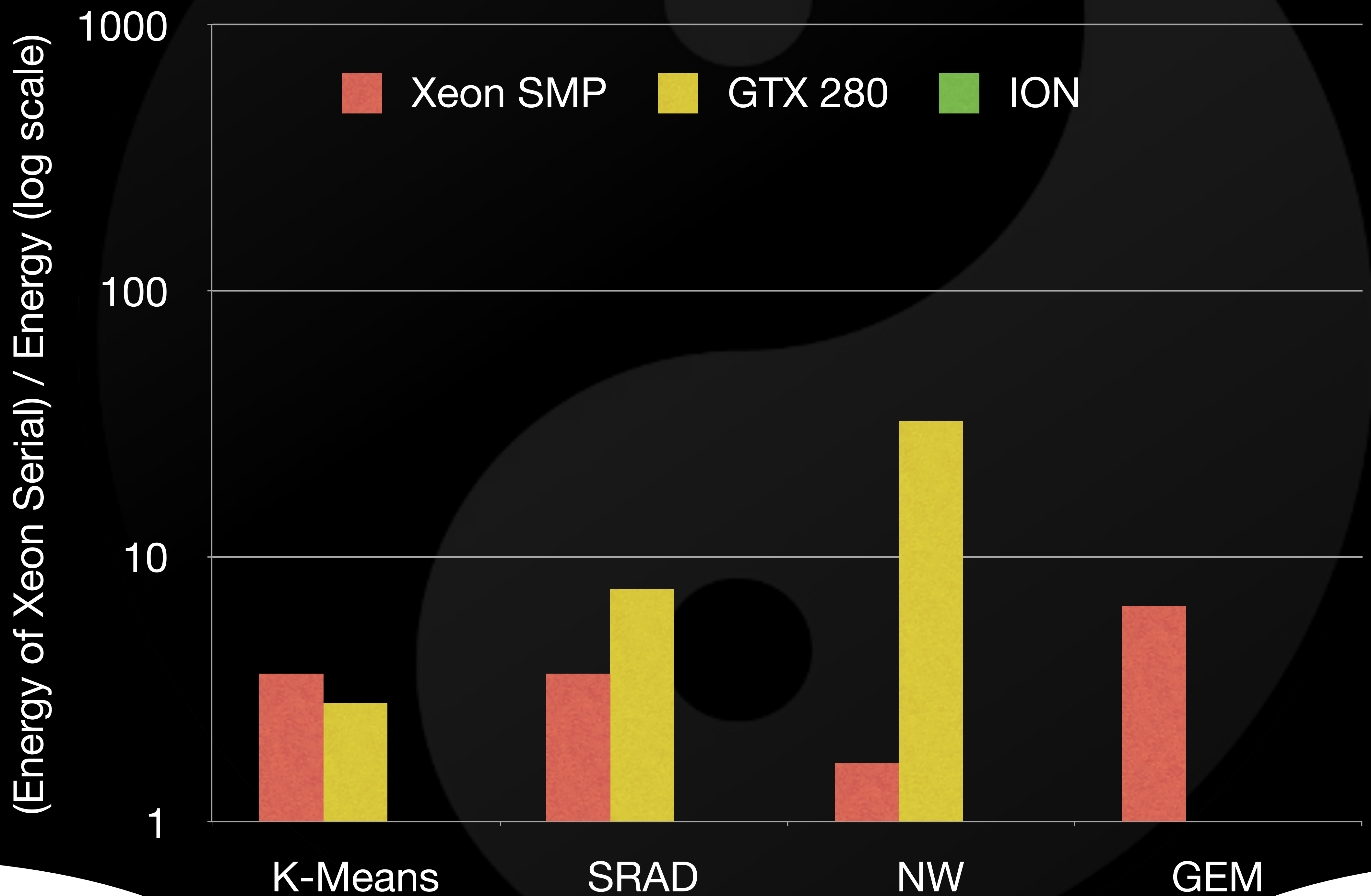


SyNeRG
http://synergy.cs.vt.edu

50

VirginiaTech
Invent the Future

Energy Overall



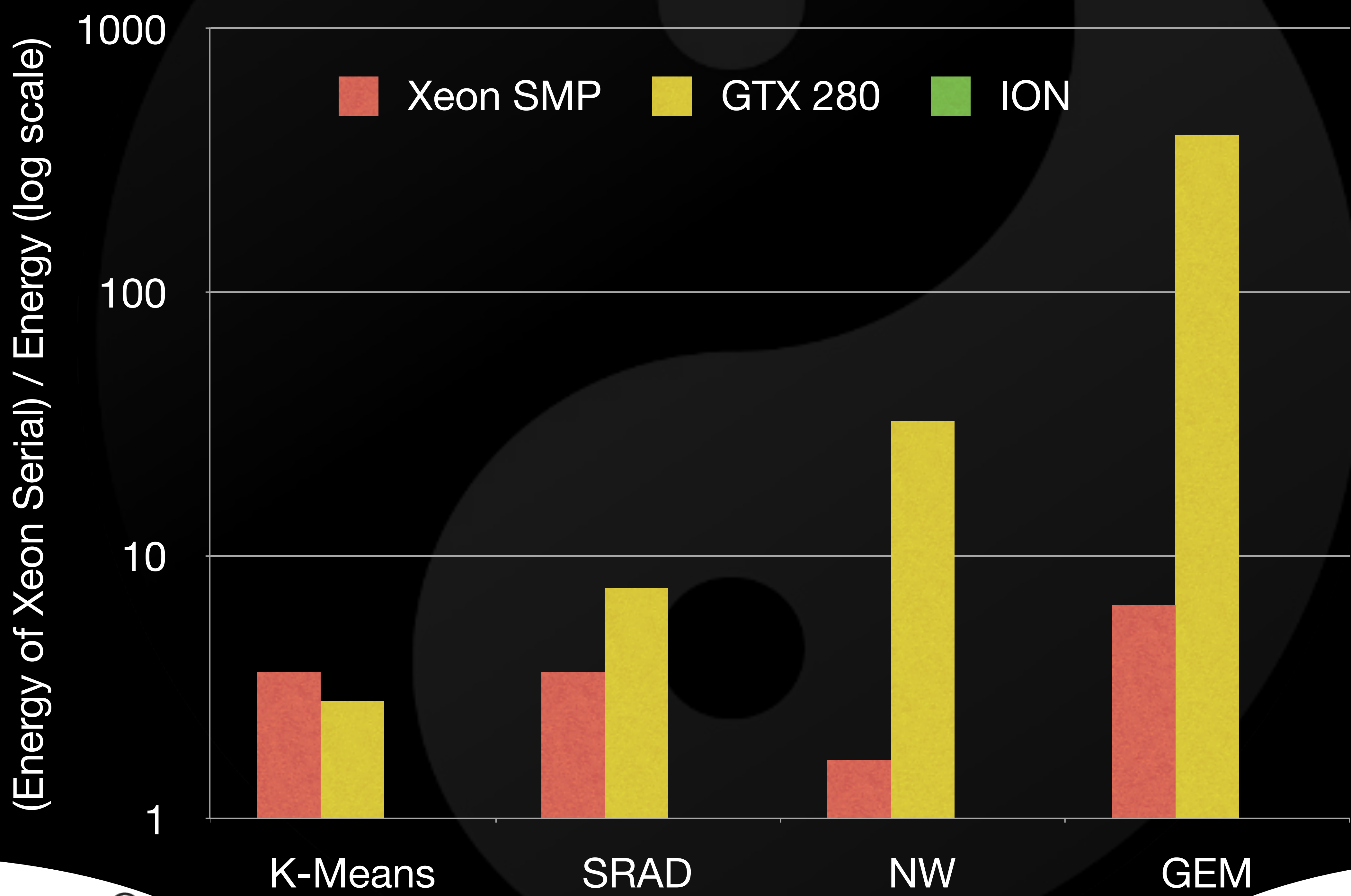
SyNeRG

http://synergy.cs.vt.edu

50

VirginiaTech
Invent the Future

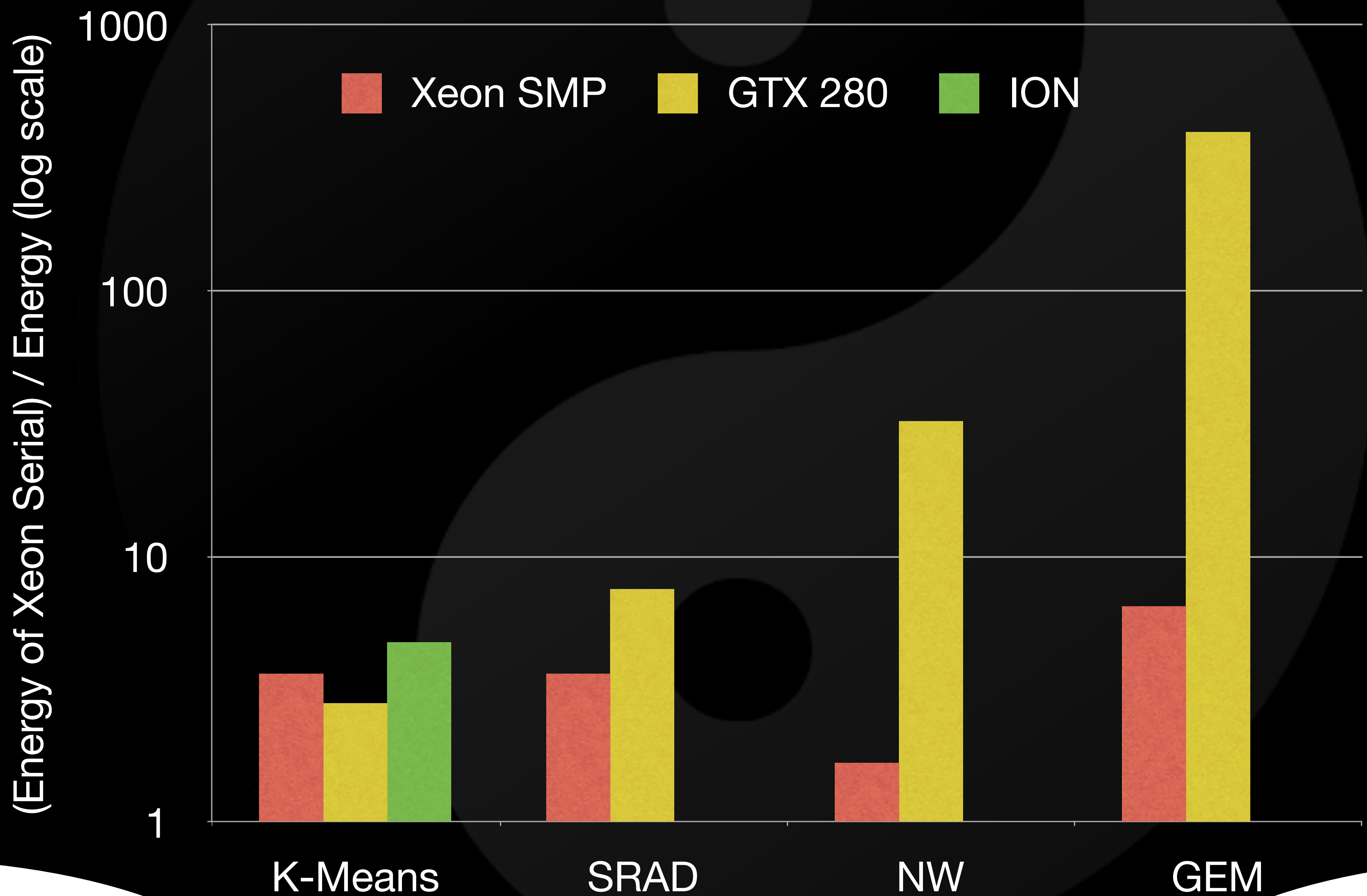
Energy Overall



SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

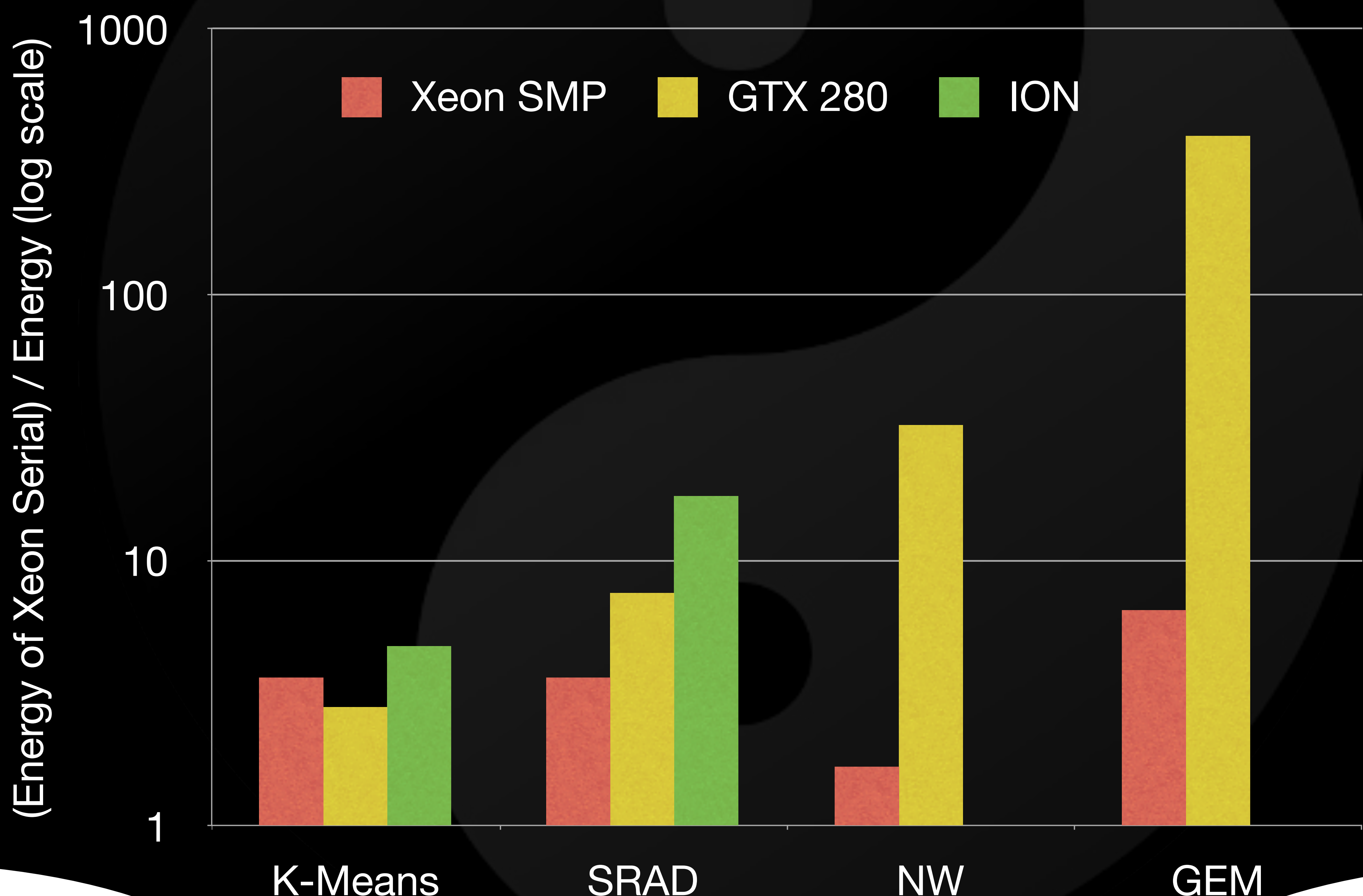
Energy Overall



SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

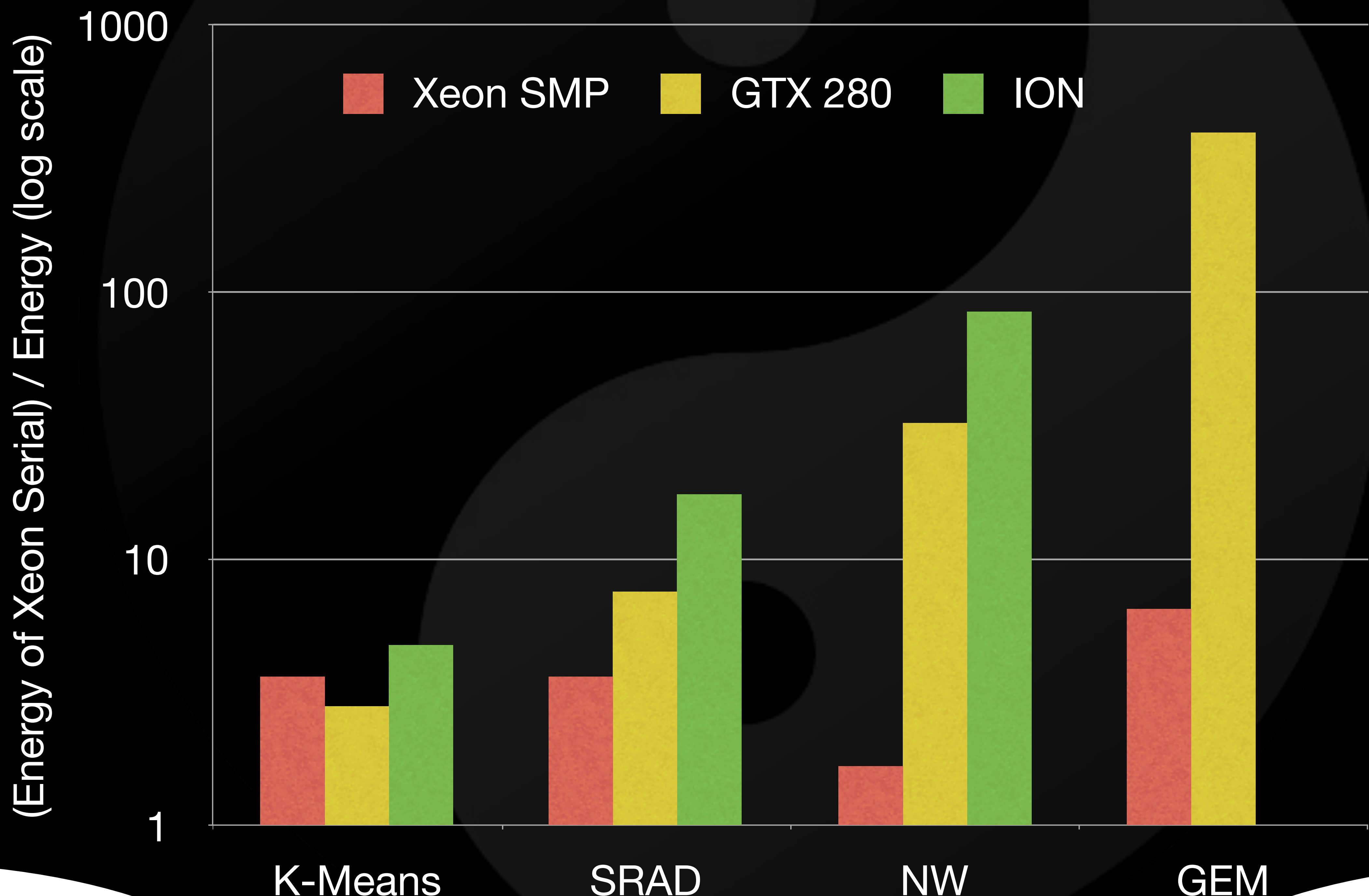
Energy Overall



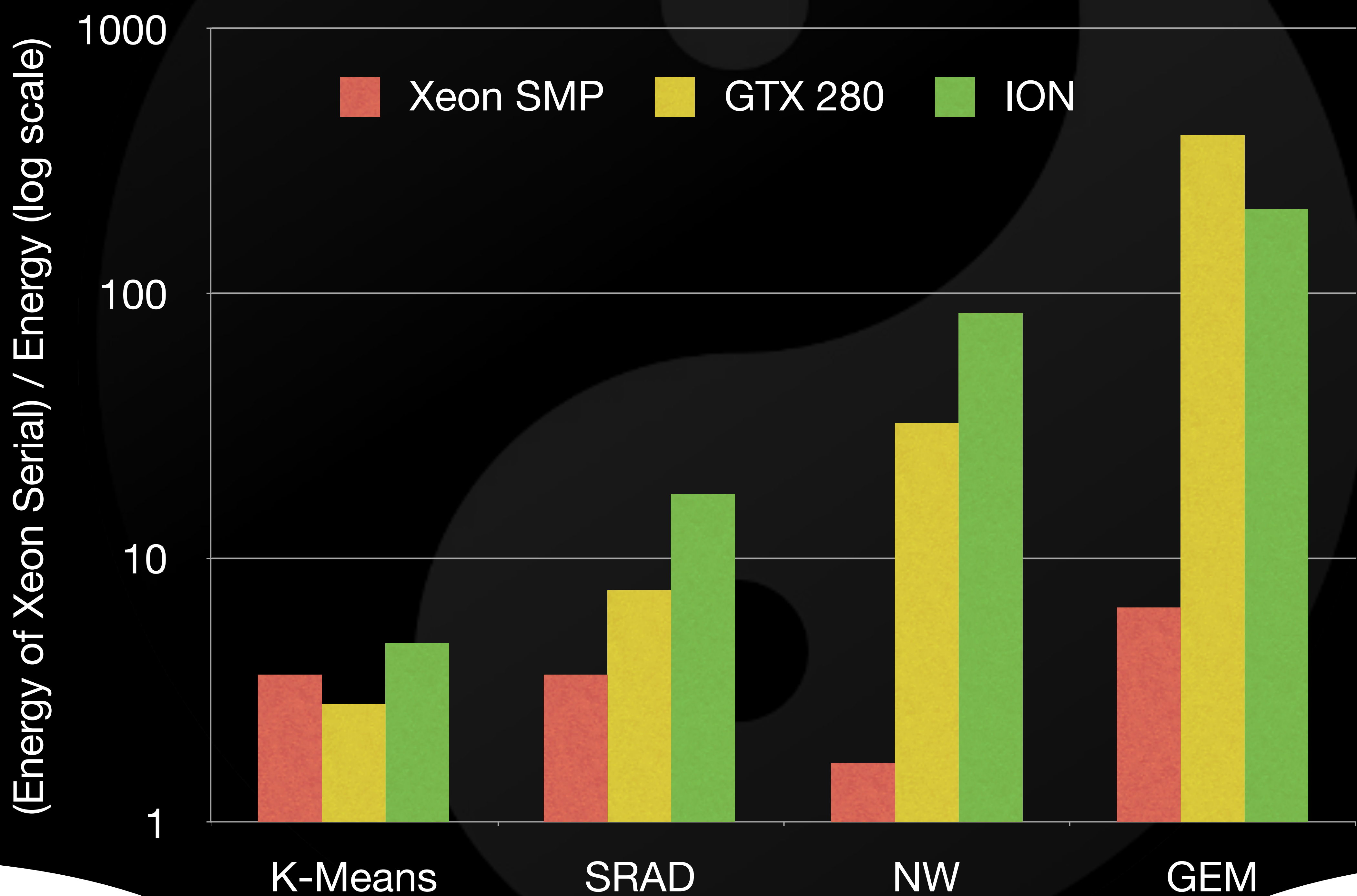
SyNeRG
http://synergy.cs.vt.edu

VirginiaTech
Invent the Future

Energy Overall



Energy Overall



Overall

	Xeon SMP	GTX 280	ION
Performance			
Power			
EDP			
Energy			

Best Average Worst

51

Conclusions

Conclusions

- GPUs are “Green” for some applications

Conclusions

- GPUs are “Green” for some applications
- Discrete GPUs are highly efficient according to EDP

Conclusions

- GPUs are "Green" for some applications
- Discrete GPUs are highly efficient according to EDP
- Integrated GPUs are highly efficient according to total energy consumption

Conclusions

- GPUs are "Green" for some applications
- Discrete GPUs are highly efficient according to EDP
- Integrated GPUs are highly efficient according to total energy consumption
- Integrated and low-power GPUs are viable options for power constrained systems

Future Work

53

Future Work

- Evaluation of further platforms:
  - AMD GPUs
  - Embedded GPUs (e.g. tegra)
  - Traditional embedded accelerators

Future Work

- Evaluation of further platforms:
  - AMD GPUs
  - Embedded GPUs (e.g. tegra)
  - Traditional embedded accelerators
- Expand into OpenCL testing

Future Work

- Evaluation of further platforms:
  - AMD GPUs
  - Embedded GPUs (e.g. tegra)
  - Traditional embedded accelerators
- Expand into OpenCL testing
- Investigate clustering low-power GPUs

Questions?

Contact

Email: tom.scogland@vt.edu

Questions?

Contact
Email: tom.scogland@vt.edu