ATKINS
Member of the SNC-Lavalin Group

SATGPU
- A Step Change in Model Runtimes
User Group Meeting

Thursday 16th November 2017

Ian Wright, Atkins

Peter Heywood, University of Sheffield

The University Of Sheffield.

SATURN

20 November 2017

1

SATGPU: Phased Development

ATKINS
Member of the SNC-Lavalin Group

Phase 1 (2016)

- Collaboration between Atkins, Highways England & TS Catapult
- Specialist resources from University of Sheffield & The Hartree Centre
- 50% / 50% private and public funding
  - demonstrate to modelling industry feasibility and value of applying GPU technology
- AND
  - provide a step change to SATURN runtimes using GPU technology

Phase 2 (2017)

- Atkins-funded investment to optimise performance
- Collaboration with University of Sheffield

The University Of Sheffield.

SATURN

Some of the Practical Challenges with Assignment

ATKINS
Member of the SNC-Lavalin Group

- Larger and more complex models
  - More zones, more detailed network, increased segmentation (hence UCs)
  - E.g. HERTMs, LoHAMs
- Wider range of scheme sizes to evaluate
- Smaller differences between scenarios
  - Requiring higher convergence targets
- Greater emphasis of Value for Money
  - More sensitivity tests & uncertainty
- Significant increase in CPU required

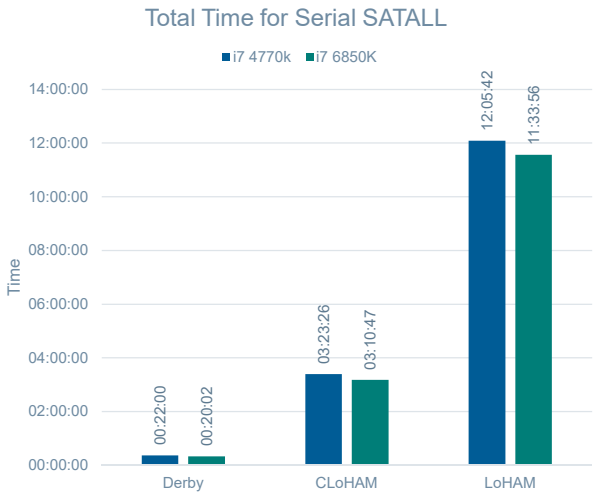
The University Of Sheffield.

SATURN

Problem: Simulation Time

ATKINS
Member of the SNC-Lavalin Group

- Large models take time to execute on CPUs
  - ~ 12 hours in serial (i7-4770k) for 5194 zones
  - Assignment can account for 97% of total runtime



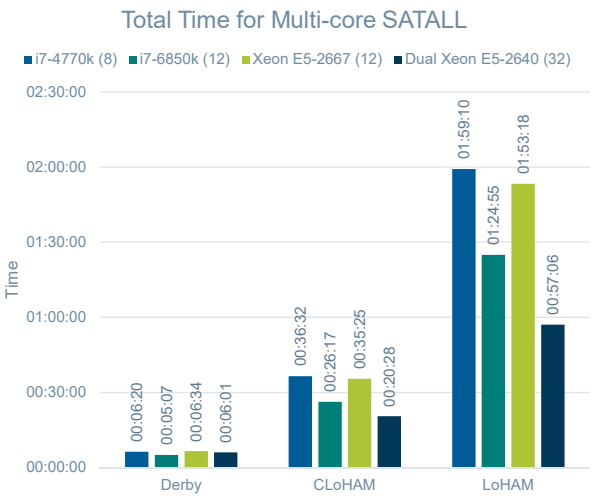
The University Of Sheffield.

SATURN

Problem: Simulation Time

ATKINS
Member of the SNC-Lavalin Group

- Large models take time to execute on CPUs
  - ~ 12 hours in serial (i7-4770k) for 5194 zones
  - Assignment can account for 97% of total runtime
  - ~ 1 hour using 2, 8 core Xeon E5-2640 CPUs (32 threads)



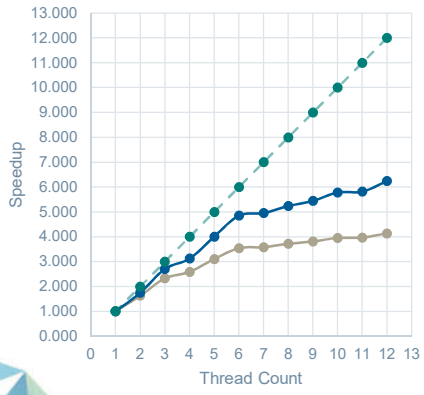
The University Of Sheffield.

SATURN

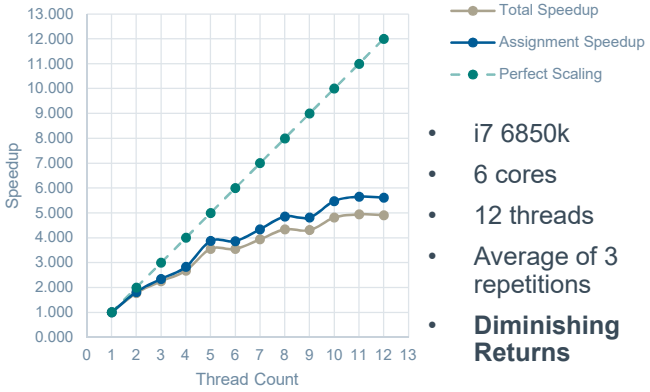
Problem: Simulation scaling

ATKINS
Member of the SNC-Lavalin Group

CLoHAM Single Assignment-Loop Speedup against Thread Count



LoHAM Single Assignment-Loop Speedup against Thread Count



- i7 6850k
- 6 cores
- 12 threads
- Average of 3 repetitions
- **Diminishing Returns**

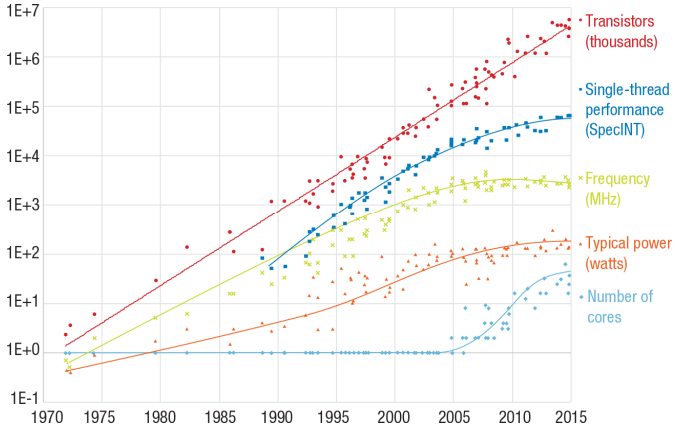
The University Of Sheffield.

SATURN

Why so slow?

ATKINS
Member of the SNC-Lavalin Group

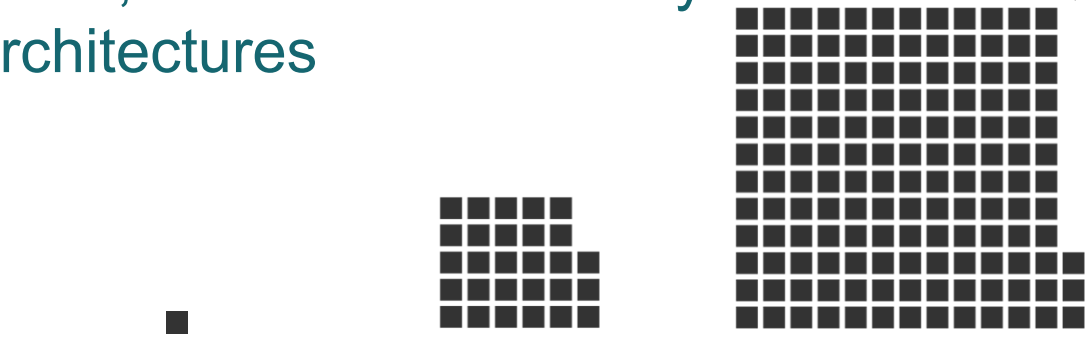
- 90s saw great improvements to CPU performance
  - 1980s to 2002: 100% performance increase every 2 years
  - 2002 to now: ~40% every 2 years
  - CPUs becoming more and more parallel



Adapting to Thrive in a New Economy of Memory Abundance, K Bresniker et al.

Serial, Multi-core and Many-core Architectures

ATKINS
Member of the SNC-Lavalin Group



Serial Computing
~53 GigaFLOPS

1 core

Parallel Computing
~1.5 TeraFLOPS

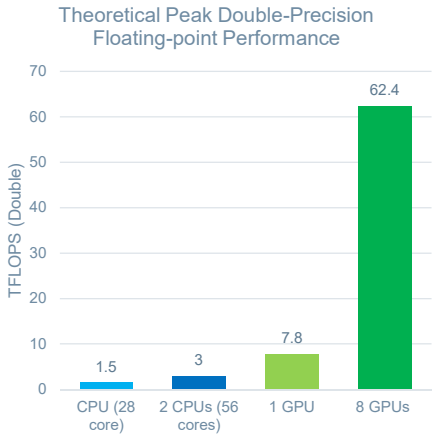
28 cores

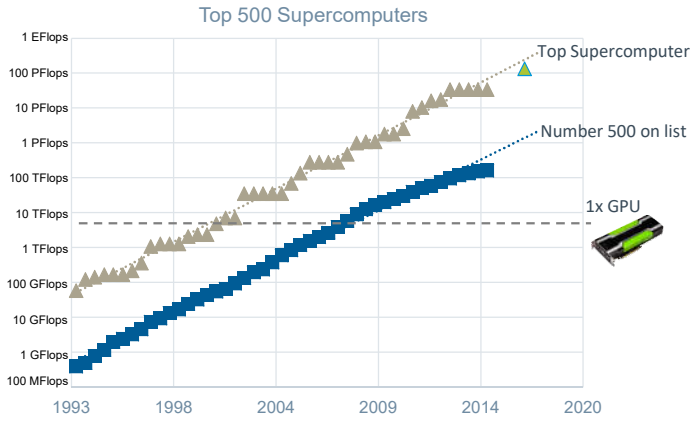
Accelerated Computing
7.8 TeraFLOPS

5120 cores

Theoretical Peak Performance

ATKINS
Member of the SNC-Lavalin Group





- Up to 8 GPUs in a single machine
- Only 2 CPUs

- 8 Volta GPUs would have been the 8th most powerful computer in the world only 10 years ago (2007)

The University Of Sheffield.

SATURN

Challenges of GPUs

ATKINS
Member of the SNC-Lavalin Group

- Using GPUs is not straightforward
  - Considerable changes required to achieve high performance
    - New algorithms (data-parallel not task-parallel)
    - New data-structures and data-management
    - Potentially re-write significant portion of application
  - Problems must expose enough parallelism to fully utilise the GPU hardware
  - Requires lots of expertise and knowledge of the hardware

GPU Computing Sheffield

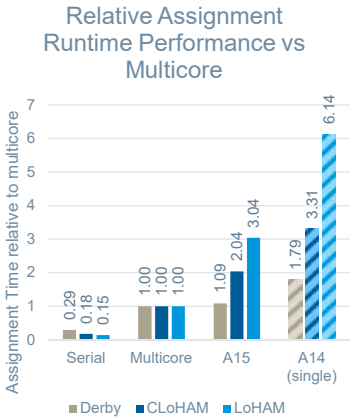
The University Of Sheffield.

SATURN

Phase 1 Technical Development

ATKINS
Member of the SNC-Lavalin Group

- Move computationally expensive path building algorithm to the GPU
  - D'Esopo-Pape replaced with Bellman-Ford
- Increase parallelism to fully utilise GPU
  - Build paths for all origin-destinations pairs concurrently (for a single user-class)
- Minimise data-transfer
  - Accumulate flow-per-link on the GPU
- Results stable but non-deterministic
  - Due to multiple equivalent-cost routes through large networks



The University Of Sheffield.

SATURN

Phase 2 Technical Development

ATKINS
Member of the SNC-Lavalin Group

Objectives

- Build upon Phase 1 addressing:
  - More consistent performance across different model sizes
  - Problem of replication
- Further development:
  - Updates to algorithm / memory (data) management
  - Optimise for Pascal architecture GPUs
  - Scaling across multiple GPUs (including load balancing)
  - Outperform dual-socket Intel Xeon Workstation (16 cores, 32 threads)
  - Update to SATURN v11.3.12W Release

The University Of Sheffield.

SATURN

Evaluation Framework - Test Hardware (i)

ATKINS
Member of the SNC-Lavalin Group

Work-station	Budget (excl. VAT)	Model	Spec	Cores / Threads
Low	£1k	HP Z820	1 x Intel Xeon E3-1270 v3 16Gb RAM, Windows 7	4C / 8T (3.0 GHz)
Medium	£2.5k	HP Z840	1 x Intel Xeon E5-2687W v4 128Gb RAM, Windows 10	12C / 24T (3.0 GHz)
High	£5k	HP Z840	2 x Intel Xeon E5-2687W v4 128Gb RAM, Windows 10	24C / 48T (3.0 GHz)

The University Of Sheffield.

SATURN

Evaluation Framework - Real-life models

ATKINS
Member of the SNC-Lavalin Group

- Representative cross-section of SATURN models
  - CLoHAM closest in size to the five RTMs

Model	Size	Zones	User Classes	Simulated Junctions	Runtimes (mins)*			
					PC £1k		PC £2.5k	PC £5k
					Single Core	Multi-Core	Multi-Core	Multi-Core
Derby	'M'	547	13	3,686	*16*	*5*	*5*	*4*
CLoHAM	'L'	2,548	5	12,932	*197*	*37*	*21*	*18*
LoHAM	'XL'	5,194	5	25,575	*659*	*119*	*62*	*50*

*Note: * Base Year AM without PASSQ*

The University Of Sheffield.

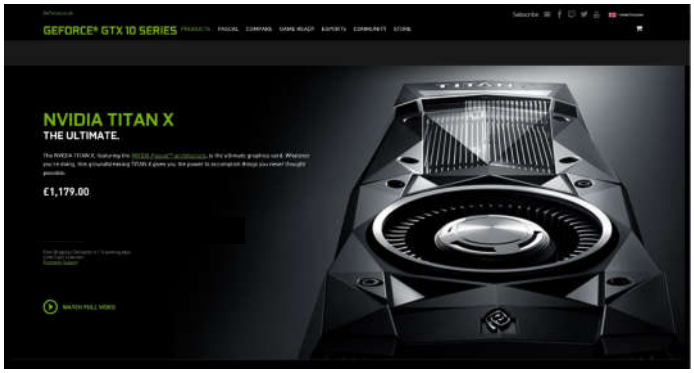
SATURN

Evaluation Framework - Test Hardware (ii)

ATKINS
Member of the SNC-Lavalin Group

Target GPU
- Nvidia Titan Xp

- Pascal Architecture
- 3840 CUDA Cores
- 12Gb RAM
- 12.5 TFLOPS (FP32)
- Price < £1k each
- Up to 3 or 4 GPUs in standard desktops
- Modes: TCC or Windows

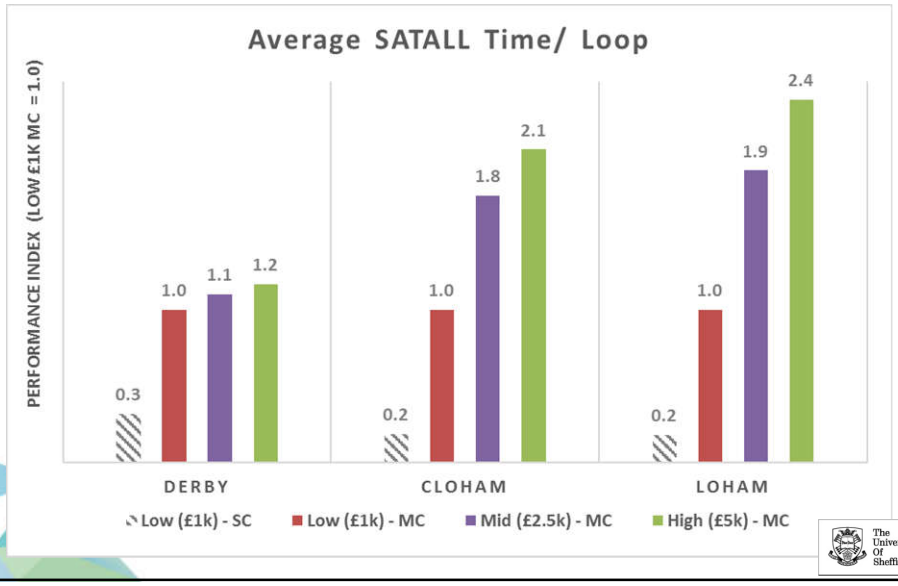


The University Of Sheffield.

SATURN

Performance: Multi-Core

ATKINS
Member of the SNC-Lavalin Group



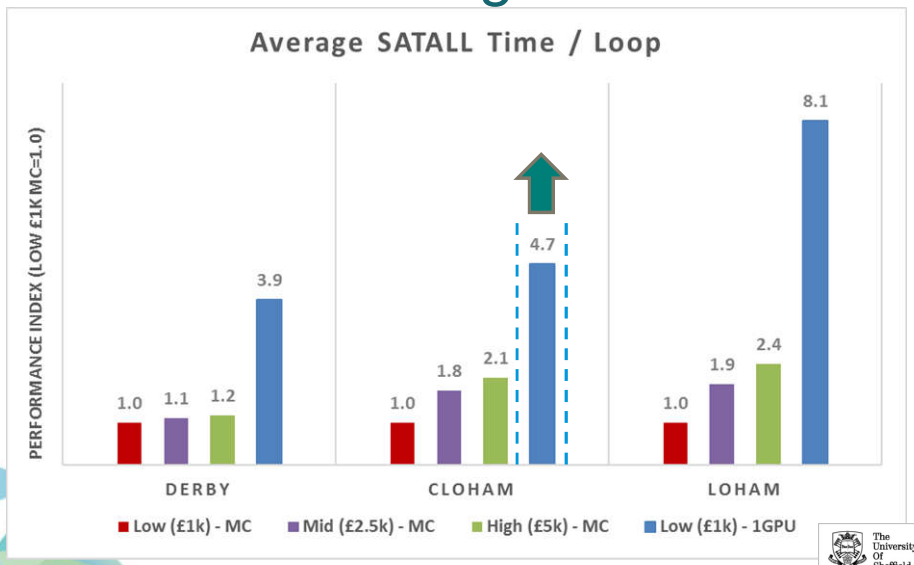
Comments:

- Extra threads = Reduced runtimes
- Assignment is Multi-threaded but Simulation remains sequential

The University Of Sheffield.

SATURN

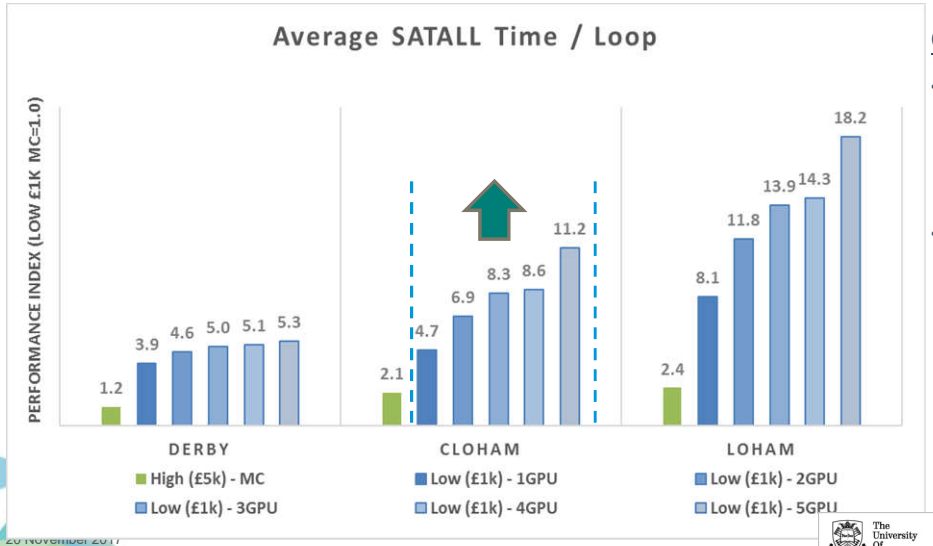
Performance: Single GPU



Comments:

- For a Low-end (£1k) desktop, adding a £1k GPU speeds-up runtimes by up to a factor of 8
- Low-end (£1k) desktop + £1k GPU is up to 4x faster than High-end (£5k) desktop

Performance: Multi-GPU

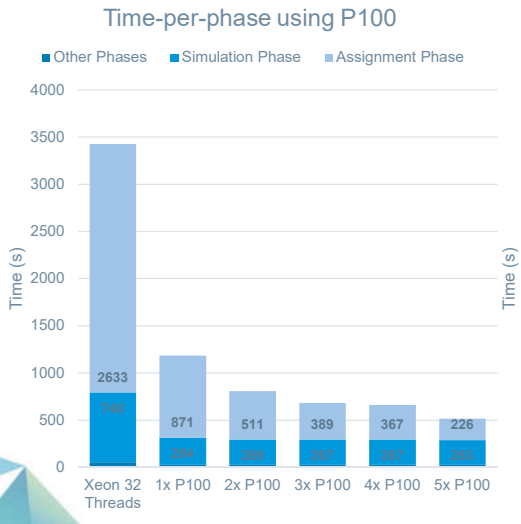


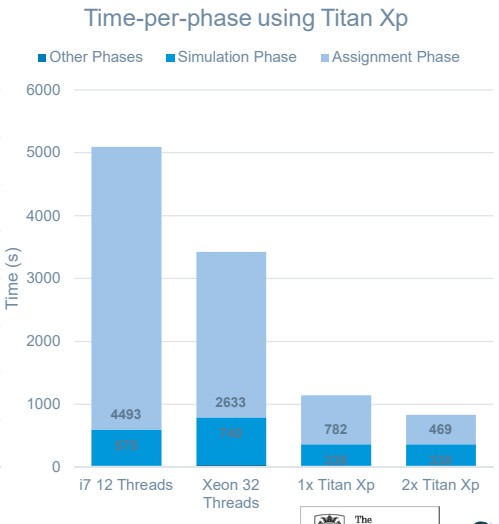
Comments:

- For a Low-end (£1k) desktop, adding a £1k GPU speeds-up runtimes by up to a factor of 8
- Load-balancing:
  - Existing method based on allocating each User Class to a specific GPU
  - Updated method will sub-dividing UCs to further boost performance for some models

Real-world Performance – LoHAM

ATKINS
Member of the SNC-Lavalin Group





The University Of Sheffield.

SATURN

Performance: Multi-GPU (ii)

ATKINS
Member of the SNC-Lavalin Group

PC	Product	Latest Test Model Runtimes (mins) */**		
		Derby	CLoHAM	LoHAM
Low (£1k)	Multi-Core	5.0	36.5	118.6
Medium (£2.5k)	Multi-Core	4.5	20.9	62.0
High (£5k)	Multi-Core	4.3	17.8	49.9
Low (£1k)	1 GPU (Titan Xp)	1.3	7.7	14.6
	2 GPUs (Titan Xp)	1.1	5.3	10.1
	3 GPUs (Titan Xp)	1.0	4.4	8.6
	And if PC has sufficient space / power			
	4 GPUs (Titan Xp)	*1.0*	*4.3*	*8.3*
	5 GPUs (Titan Xp)	*0.9*	*3.3*	*6.5*

5.8x faster

7.7x faster

 *Note: * Base Year AM without PASSQ; ** **Dependent on model / hardware / software settings***

Performance: Convergence & Summary Statistics

ATKINS
Member of the SNC-Lavalin Group

- Reported for LoHAM – similar differences for Derby & CLoHAM

Measure / Criterion		SM	B8	%Diff
Convergence*	Stability (%Flows)	97.6%	97.5%	
	Stability (%Delays)	98.9%	98.6%	
	Proximity (%GAP)	0.013%	0.019%	
Summary Statistics	Total Distance (pcu-km)	115,782,720	115,783,080	0.00%
	Total Time (pcu-hrs)	1,727,833	1,728,068	+0.01%
	Total Delay (pcu-hrs)	158,058	158,300	-0.15%

Converged & Little Change

** Based on WebTAG M3-1 Table 4*

The University Of Sheffield.

SATURN

LoHAM: Checks on Validation

ATKINS
Member of the SNC-Lavalin Group

Measure / Criterion	Aspiration	SM	B8	Diff
Links - GEH <5	85%	64%	64%	0%
Links - GEH <7.5	85%	78%	78%	0%
Links - DMRB Flow Criteria	85%	74%	74%	0%
Screenline - Flow Difference <5%	85%	90%	89%	0%
Enclosure - Flow Difference <5%	85%	94%	94%	0%
Mini screenline - GEH <5	85%	91%	91%	0%
JT Routes - Time Difference < 15%	85%	92.1%	92.1%	0%
Links - GEH <5	85%	64%	64%	0%

No Change

The University Of Sheffield.

SATURN

Straightforward to upgrade …

ATKINS
Member of the SNC-Lavalin Group

SatGPU
User Licence

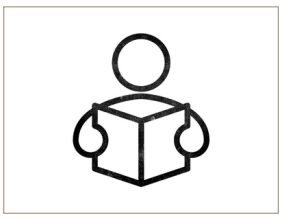
SoftwareKey

The University Of Sheffield.

SATURN

… and please note

ATKINS
Member of the SNC-Lavalin Group



© Richard de Ruijter, Dribbble.com (2013)

1. Initial release builds upon SATURN v11.3.12W release
   - For SPIDER only
2. Developed in Nvidia's CUDA for their GPUs - no AMD Radeon
3. Mix and match GPU brands, models, architecture & types
   - Maxwell(?) (2016), Pascal (2017), Volta (2018)
   - Testing with: Pascal-based Titan Xp or Geforce 1080Ti
   - Recommend Pascal
4. Small numerical differences between Single Core, Multi-Core & SatGPU
5. A single assignment fully utilises all the GPU hardware – undertake runs sequentially
6. Maximum number of Multi-Core threads is 32 across all SATALL-based assignments

The University Of Sheffield.

SATURN

Licencing and Support

ATKINS
Member of the SNC-Lavalin Group

Distribution

- Single User Add-on
- Controlled using Digital licence
  - Uses Solo Software (as per SatView)
- Available for Levels N3, N4 & X7

Pricing

- FY17/18 Purchase Price
  - Up to £11k per single user licence
  - Discount for additional licences
- Annual Support Fee
  - 20% of Prevailing Sales Price

FY17/18 Price – Early Adopter Pricing

Examples	Level G4 (N4)		Level G7 (X7)	
	Per Licence	Annual Support	Per Licence	Annual Support
Licence 1	£9,900	£1,980	£10,500	£2,100
Licences 2+	P.O.A.		P.O.A.	
Single Licence + FY17/18 Support	£10,395		£11,025	

Upgrade route for new NVidia GPUs

- Volta due in summer 2019
- Estimated performance increase of up to 30%
- Upgrade option to unlock for 'Next Gen' GPU Hardware

The University Of Sheffield.

SATURN

Work Items & Timescales

ATKINS
Member of the SNC-Lavalin Group

In progress:

- Improved load-balancing across multiple GPUs by splitting UCs
- Secondary analysis using UFOs
- Digital Licence
- Final benchmarking on wider range of hardware

Pending:

- Benchmarking across more models (e.g. HERTMs)
- Commercial release – early Jan'18

The University Of Sheffield.

SATURN