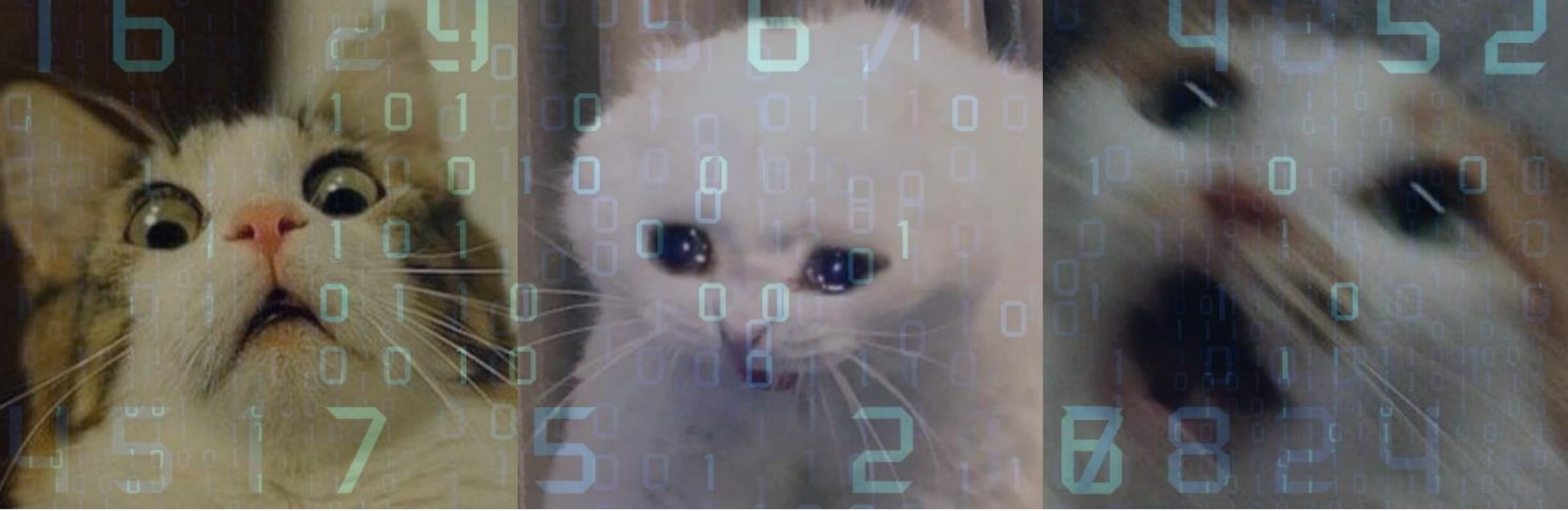


Don’t let the reality of GC×GC-MS data burst your bubble! Or how the &@$#%* am I supposed to manage all these bits and bytes?!?

James J. Harynuk,[1] Broderick Wood[2]

1- Department of Chemistry, University of Alberta / TMIC, Edmonton, Canada

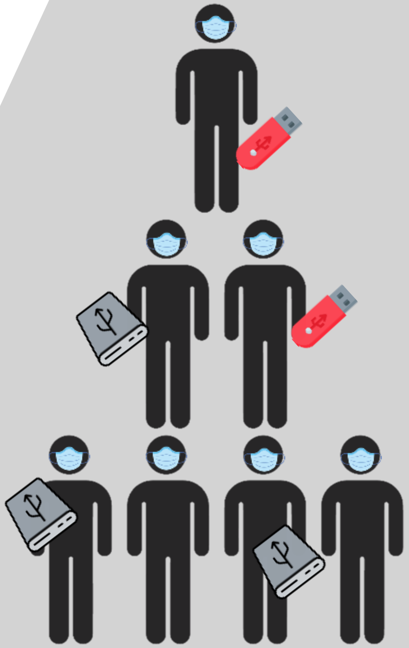
2- IST, University of Alberta

15th Multidimensional Chromatography Workshop

Challenges for GC×GC labs

- How to protect data from loss?
- How to move data from place to place?
- What should I get for a data processing computer?
  - Vendors specify “minimum requirements”
  - Vendors sending computers that cannot handle data.
  - Where should I spend $$$???

Us a few years ago…



Data Acquisition

ChromaTOF

ChromSpace

Data Processing

Then we got a big grant

- More instruments
- More students
- More headaches…

INNOVATION.CA

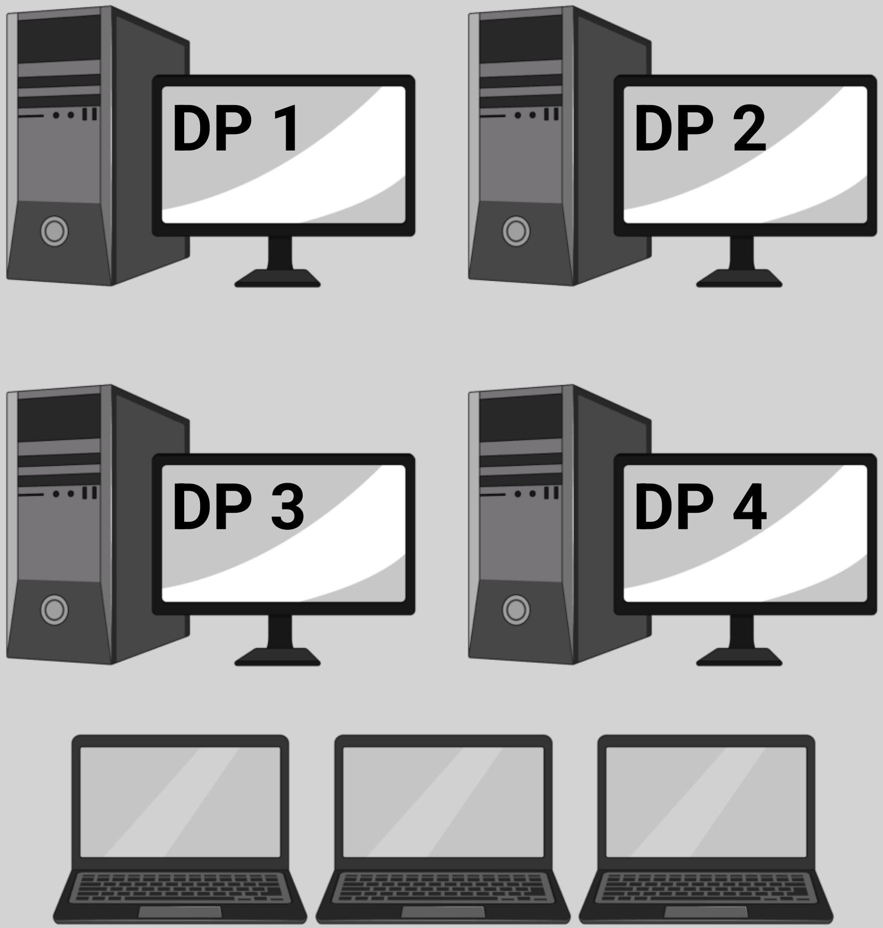
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION

Us now..

Data Acquisition

ChromaTOF®

Data Processing



New situation

- Data from 3 instruments, legacy data from another
- Many students
- Many clients
- Four main data processing machines

- Questions
  - How to move data efficiently and protect it?
  - Where to spend money on new data processing machines?

Two types of lab

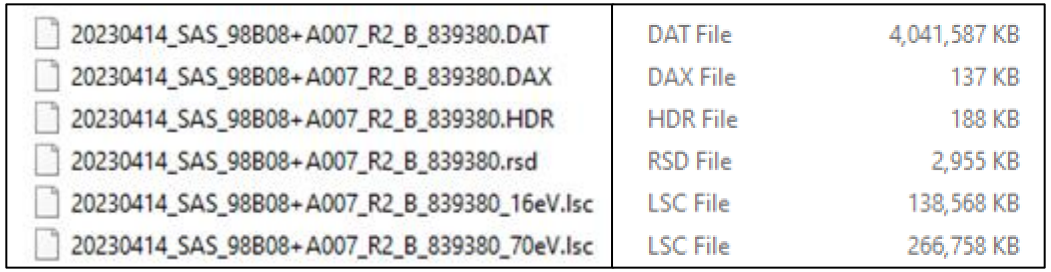
SMALL LAB	BIG LAB
• 1 instrument	• 2+ instruments
• 1-2 DP machine(s)	• 2+ DP machines
• 1-2 users	• Many users / projects

Data storage/management goals

Data should…

- move off of instrument CPUs automatically
- be stored/backed up immediately
- be accessible to users
- be safe from users
- be safe from the outside world

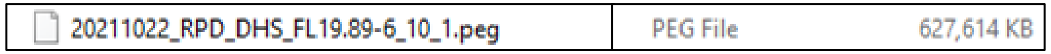
How much space do you need?



20230414_SAS_98B08+A007_R2_B_839380.DAT	DAT File	4,041,587 KB
20230414_SAS_98B08+A007_R2_B_839380.DAX	DAX File	137 KB
20230414_SAS_98B08+A007_R2_B_839380.HDR	HDR File	188 KB
20230414_SAS_98B08+A007_R2_B_839380.rsd	RSD File	2,955 KB
20230414_SAS_98B08+A007_R2_B_839380_16eV.lsc	LSC File	138,568 KB
20230414_SAS_98B08+A007_R2_B_839380_70eV.lsc	LSC File	266,758 KB

1h GC×GC run, 100 Hz
40-600 m/z range

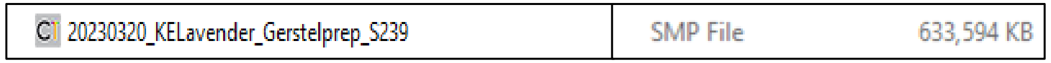
Tandem EI on BenchTOF
~ 4,500 MB per sample and 6 files



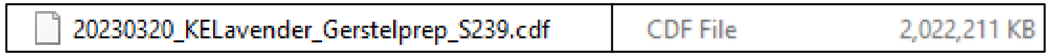
20211022_RPD_DHS_FL19.89-6_10_1.peg	PEG File	627,614 KB

Pegasus IV (ChromaTOF 4.x) .peg file
~625 MB per sample, one file

1h GC×GC run, 200 Hz
40-500 m/z range

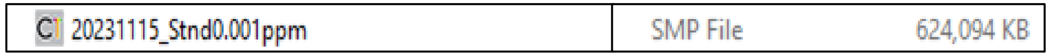


20230320_KELavender_Gerstelprep_S239	SMP File	633,594 KB



20230320_KELavender_Gerstelprep_S239.cdf	CDF File	2,022,211 KB

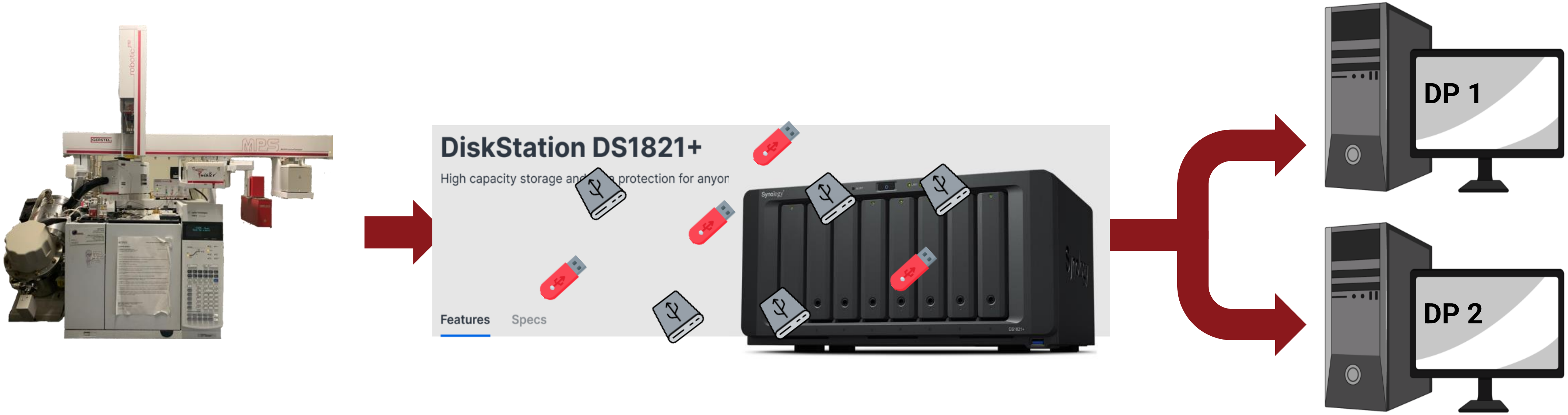
Pegasus BT SMP file
~630 MB per sample, one file
~2,200 MB (dumped out as .CDF file)



20231115_Stnd0.001ppm	SMP File	624,094 KB

Pegasus HRT+ SMP file
~625 MB per sample, one file

In a smaller lab…

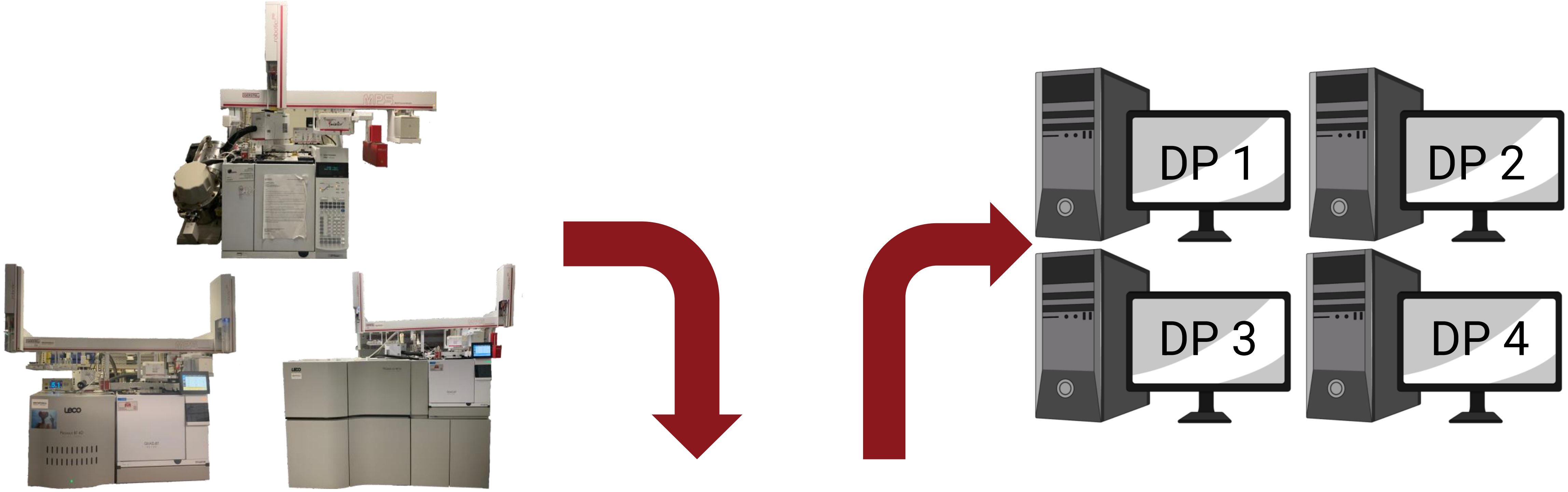


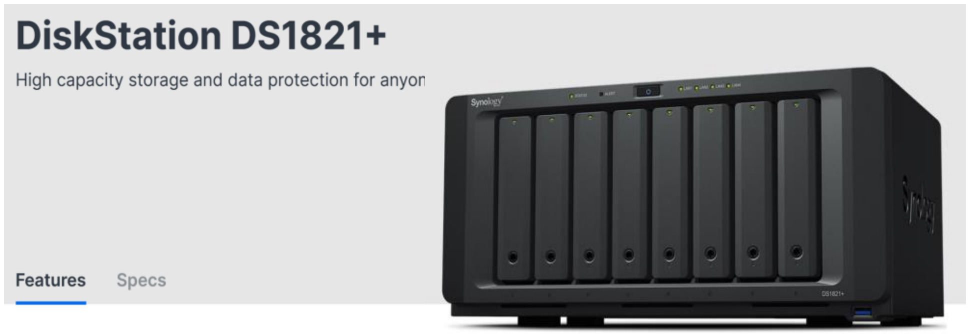
Network storage is relatively inexpensive and highly effective

- Synology DiskStation 1821+ is ~$1300CAD
- Stack of disk drives with small computer, RAM buffer (4GB) and network card
- With stack of ***n*** drives in RAID 5 array…
  - (***n-1***)× Storage (e.g. 8× 8TB drives ≈ 56 TB space)
  - ~ ***n***× write speed (write speed of 7200 RPM HDD ~80-150 MB/s)
  - Data is safe if a drive fails

Better to use 8× 4TB than 4× 8TB

In a bigger lab…





DiskStation DS1821+

High capacity storage and data protection for anyon

What if I need more space?

This model easily expands by 5 or 10 drives

8-bay system $1300
20 TB Ironwolf drive $480

$5000 CAD for 140 TB

What about data processing??

Computer

Where to put money?

- CPU?
- RAM?
- GPU?

Software for processing???

Vendor’s Software?

3rd party software?

How to dump to .cdf, matlab, etc…

Software for Processing

Small studies (Pegasus IV)

ChromaTOF®

- ChromaTOF 4.x with stat compare
  - Can be slow if not careful managing drive space

Big data sets (Pegasus IV) + all BenchTOF

GC Image
Software for Multidimensional Chromatography

- GCImage
  - Fast, reasonably robust alignment

Desired output is aligned peak table

- Freedom to use any chemometric / ML tools we want

Software for Processing

Data from BT/HRT

ChromaTOF 5.x

- Peak picking is much improved
- Alignment / fusion across samples still needed

For large studies and new processing tools

- Dump raw data to CDF
- Pull into GCImage
- Convert to a really efficient in-house file format

What about computers?

More help from vendors would be nice…

Resource Monitor / Task Manager are your friends

- Have these open while you’re trying to work
- Which resources are causing bottlenecks

Pay close attention to process monitors / logs while processing is going

- Can point to specific steps that are slow
- Uptime matters!!!

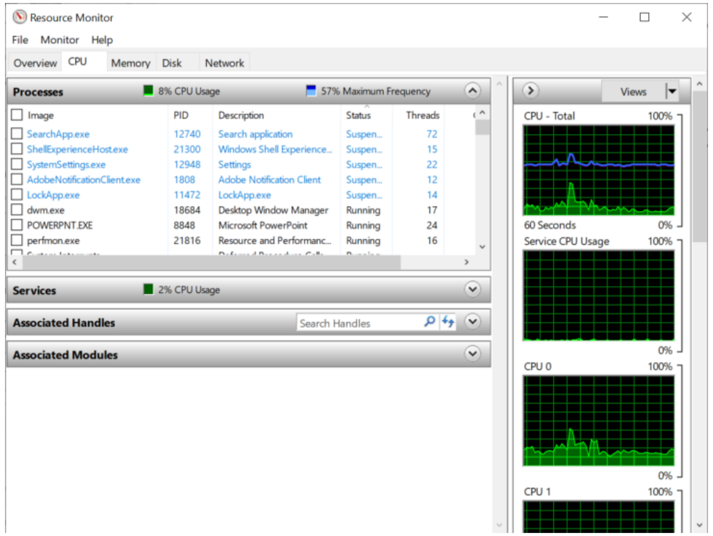
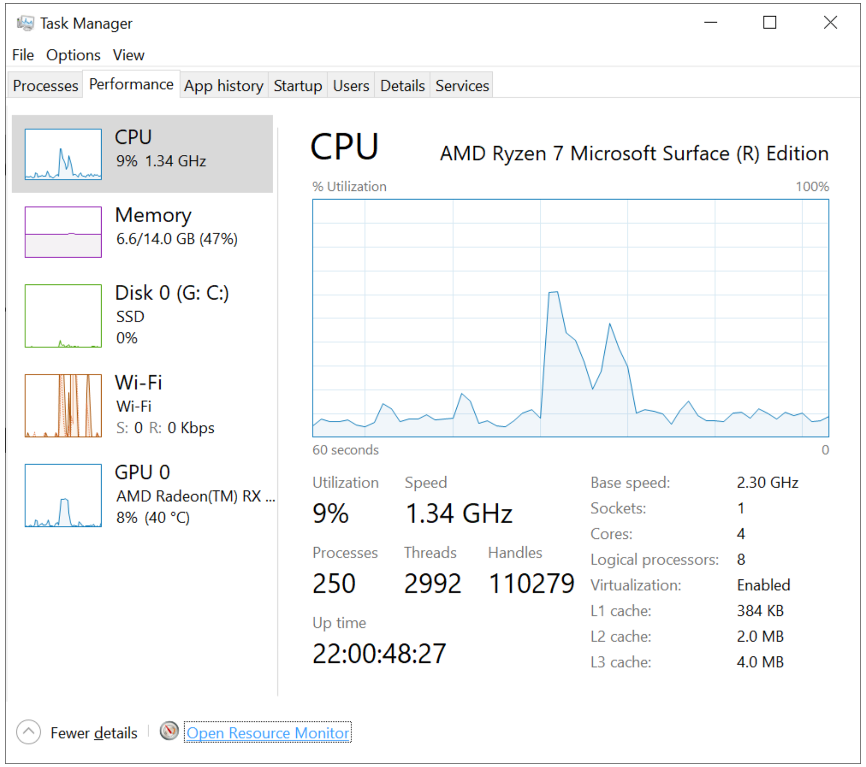


	Image	PID	Description	Status	Threads
□	SearchApp.exe	12740	Search application	Suspen...	72
□	ShellExperienceHost.exe	21300	Windows Shell Experience...	Suspen...	15
□	SystemSettings.exe	12948	Settings	Suspen...	22
□	AdobeNotificationClient.exe	1808	Adobe Notification Client	Suspen...	12
□	LockApp.exe	11472	LockApp.exe	Suspen...	14
□	dwm.exe	18684	Desktop Window Manager	Running	17
□	POWERPNT.EXE	8848	Microsoft PowerPoint	Running	24
□	perfmon.exe	21816	Resource and Performanc...	Running	16



Utilization 9%

Speed 1.34 GHz

Processes 250

Threads 2992

Handles 110279

Up time 22:00:48:27

Base speed:	2.30 GHz
Sockets:	1
Cores:	4
Logical processors:	8
Virtualization:	Enabled
L1 cache:	384 KB
L2 cache:	2.0 MB
L3 cache:	4.0 MB

What about computers?

CPU – get the best you can afford

- Had good experience with AMD Threadripper and intel i9 chips

Hard drives matter and can matter a LOT!!!

- Large 7200 RPM HDD (150 MB/s)
- Networked storage (8× 150 = 1.2 GB/s with 10 Gigabit network)
- Local RAID 0 array of 4×4 TB m2.NVMe SSD (~10-20 GB/s)

Many computers have smallish C:\ and separate D:\ for data

- Don’t leave big files in “downloads” “desktop” (these are on C:\)

What about RAM and GPU?

RAM

To get max performance out of CPU make sure every memory slot is filled

- Our AMD system has 8×16 GB; 4×32 or 2×64 are bad ideas
- In our testing never get beyond ~28 GB RAM in use

GPU

GPU is used to render graphics, drive monitors

- GPU does not seem to be relied on to process data…
- Missed opportunity, but…

Impact of drive speed

Opening a ChromaTOF 5 (BT) database with 10 samples

Local 7200 RPM HDD	**3.8 s**
Networked storage	**2.9 s**
Local NVMe RAID array	**1.6 s**

Read/write files with GC Image

Read/write speed impacting batch processing a series of .CDF or .PEG files

Local 7200 RPM HDD	**6 min/sample (14 days for 3800 samples)**
Networked storage	**1.5 min/sample (4 days for 3800 samples)**
NVMe RAID	**0.9 min/sample (2.3 days for 3800 samples)**

Conclusions

- You can set up effective tools to move/manage data
- You need to think a bit about your processing computer
  - Not hard to identify bottlenecks
- Vendors could probably push more math to GPUs

Acknowledgements

INNOVATION.CA

CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION

GenomeCanada

NSERC CRSNG

UNIVERSITY OF ALBERTA