Mor Consulting

Applied Parallel Computing with Python – Lessons Learned

PyCon 2013

Ian@MorConsulting.com @IanOzsvald - PyCon 2013

Goal

- Scalable, robust systems in the face of increasing unreliability
- Reporting and debugging to quickly fix problems
- Parallelising CPU-bound and disk-bound tasks
- Visualisations for communication

Taught before

- CPU-bound profile(runsnake, line_profiler)
- CPython objects & numpy
- Compilation (cython, shedskin)
- Efficient memory access (numexpr)
- Multi-core (multiprocessing)
- Multi-machine (pp, iPython Cluster, PiCloud)
- CUDA

About me (Ian Ozsvald)

Mor Consulting

- Data Science consultant for 14 years
- Python for 9 years (C, NLProc.)
- SocialTies and Headroid at EuroPythons
- StartupChile computer vision
- Annotate.io NLP on social media
- ShowMeDo.com co-founder
- IanOzsvald.com - MorConsulting.com

Group photo

- I'd like to take a photo - please smile :-)

Scalability

- You won't specify it correctly
- It will break
- Separability – for scaling and testing
- Vertical and Horizontal
- Bottlenecks? CPU/Disk/Mem/Network
- Assume cluster size will change (best – during production)

Coding

- VirtualBox/Vagrant – match deploy env
- Design for e.g. dev/test/staging/prod envs
  - Enable upgrade/refactor testing
- Tuples bad, dicts ok, classes better
- Use JSON for persistence

Robustness

- Assume failures occur
- Assume capacity constraints
- Test driven development
- Specify what's required in each system
- “Notes on Distributed Systems for Young Bloods” http://www.somethingsimilar.com/2013/01/14/n

Queueing approaches

- Random queue choice dangerous
- Must not flood queues
  - Use timeouts, retries
  - Check capacities

Reporting

- Server Density
- StatHat
- Greylog central logger

Tool choices

- Gael's joblib
- Avoid NIH - Celery
- Consider range of errors that can occur (e.g. Connection dropped, 500 int error, unknown URL)
- Glances/htop, dsniff, lsof, iftop, netstat
- supervisord, circus, upstart
- fabric, puppet