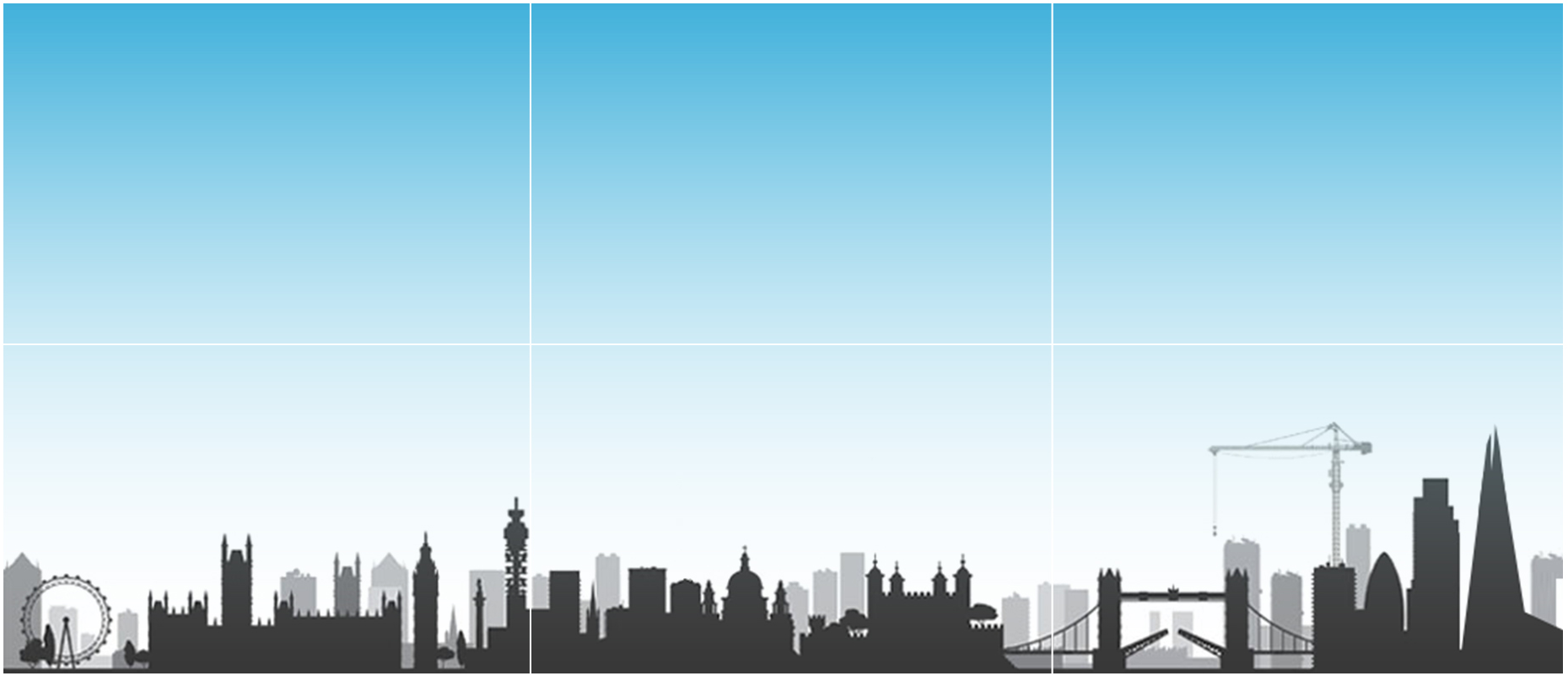
EARL 2014

Cloud Computing, Grid Computing and Docker



Mark Sellors – Mango Solutions

ABOUT ME…

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

ENOUGH ABOUT ME

EARL 2014

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

On to the interesting stuff…

- Cloud Computing
- Grid Computing
- Docker

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

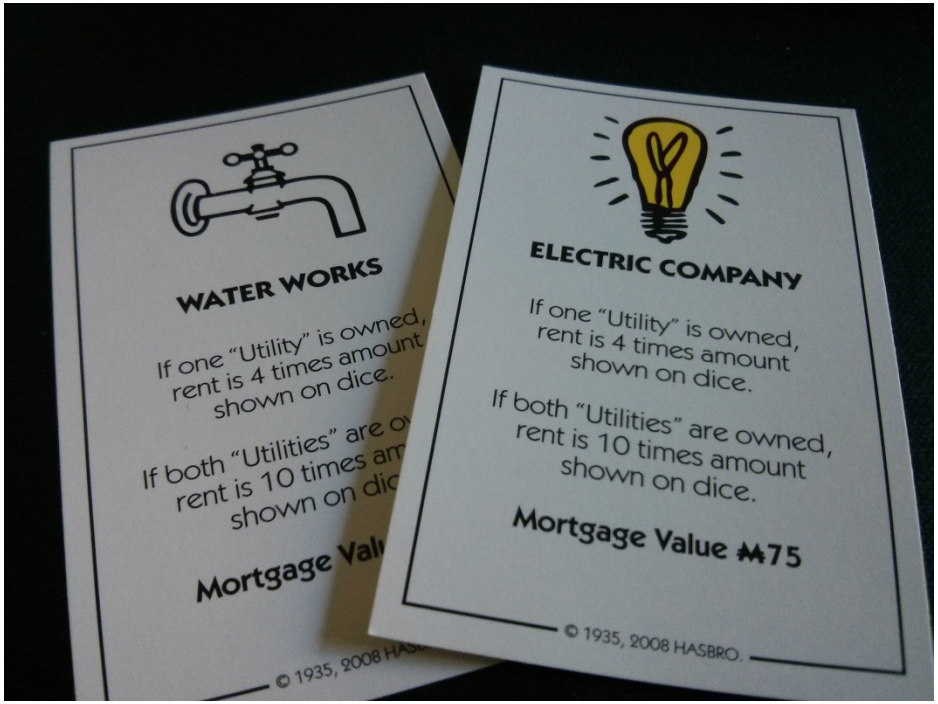
Cloud Computing

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

What is cloud computing?

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

What is cloud computing?



Utility computing

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Three main types

- SaaS
- PaaS
- IaaS

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Three main types

- SaaS – Software as a Service
  - Gmail, Dropbox, iCloud, Twitter, Facebook
- PaaS

- IaaS

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Three main types

- SaaS – Software as a Service
  - Gmail, Dropbox, iCloud, Twitter, Facebook
- PaaS – Platform as a Service
  - Heroku, Google AppEngine, Salesforce
- IaaS

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Three main types

- SaaS – Software as a Service
  - Gmail, Dropbox, iCloud, Twitter, Facebook
- PaaS – Platform as a Service
  - Heroku, Google AppEngine, Salesforce
- IaaS – Infrastructure as a service
  - Amazon AWS, Google Compute Engine, MS Azure

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

IaaS

- Of the three main type of cloud infrastructure is by far the most flexible and the one we're seeing the most use of in industry.
- Because you're essentially renting servers you then have the flexibility to do what you want with them without needing a massive IT team to support you.
- Frees analysts to work on analysis, rather than having to wait around

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Why should I care?

- Scale
- Demand
- Flexibility
- Speed

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Sounds great, what about an example?

- Problem: Machine for data analysis using R, should have ~32G RAM.
- Solution: Use Amazon Web Services to spin up a node with 30G RAM and 8 CPU's. Internally developed scripts deployed the environment in under 10 minutes.
- Total cost: $12.30

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The second example

- Problem: Our cloud based database use changes and performance is now very poor.
- Solution: Quickly migrate database to a system optimised for performance

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The downsides

- Security
- Transfer speeds
- Post pay billing
- Existing providers wont necessarily be able to help you much, especially if you're going off-piste and doing stuff that isn't officially baked into their

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The future

- Market will continue to grow
- Expect to see more niche players enter the market – For example a public analytics cloud
- Also expect to see more niche services spring up between you and the cloud.
- PAYG billing – using the same model as mobile phone providers use.

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Grid Computing

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

What is it?

- A grid is a batch processing system.
  - It takes jobs submitted by a collection of users and schedules and then runs those jobs.
- Grids can vary in size from a few processing cores to many thousands.
- Use cases.
  - Animation/SFX rendering
  - Climate modelling
  - Clinical pharmacology
  - Anywhere where large scale compute resources can help solve complex problems
- Batch processing and MPI (Message Passing Interface) form the basis of all modern supercomputers.

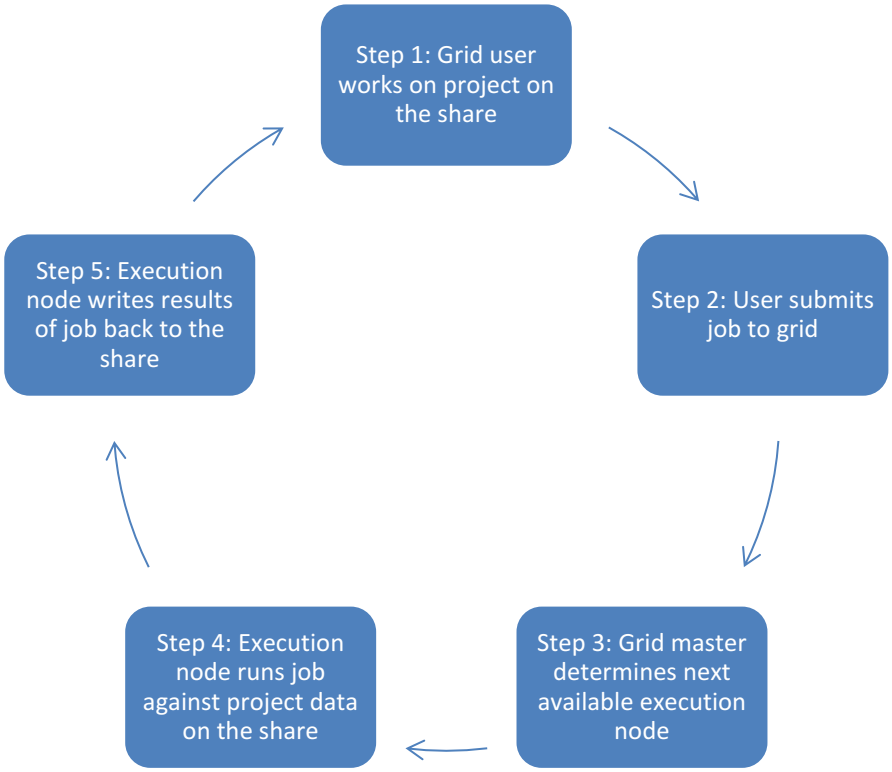
Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Some grid computing software

- SGE – Sun Grid Engine
- OGE – Oracle grid engine
- OGS – Open Grid Scheduler
- Univa – Grid Engine
- LSF – IBM/Platform Load Sharing Facility
- OpenLava – A fork of an older version of LSF
- Globus Toolkit
- gLite
- HTCondor/Condor
- PBS – Portable Batch System

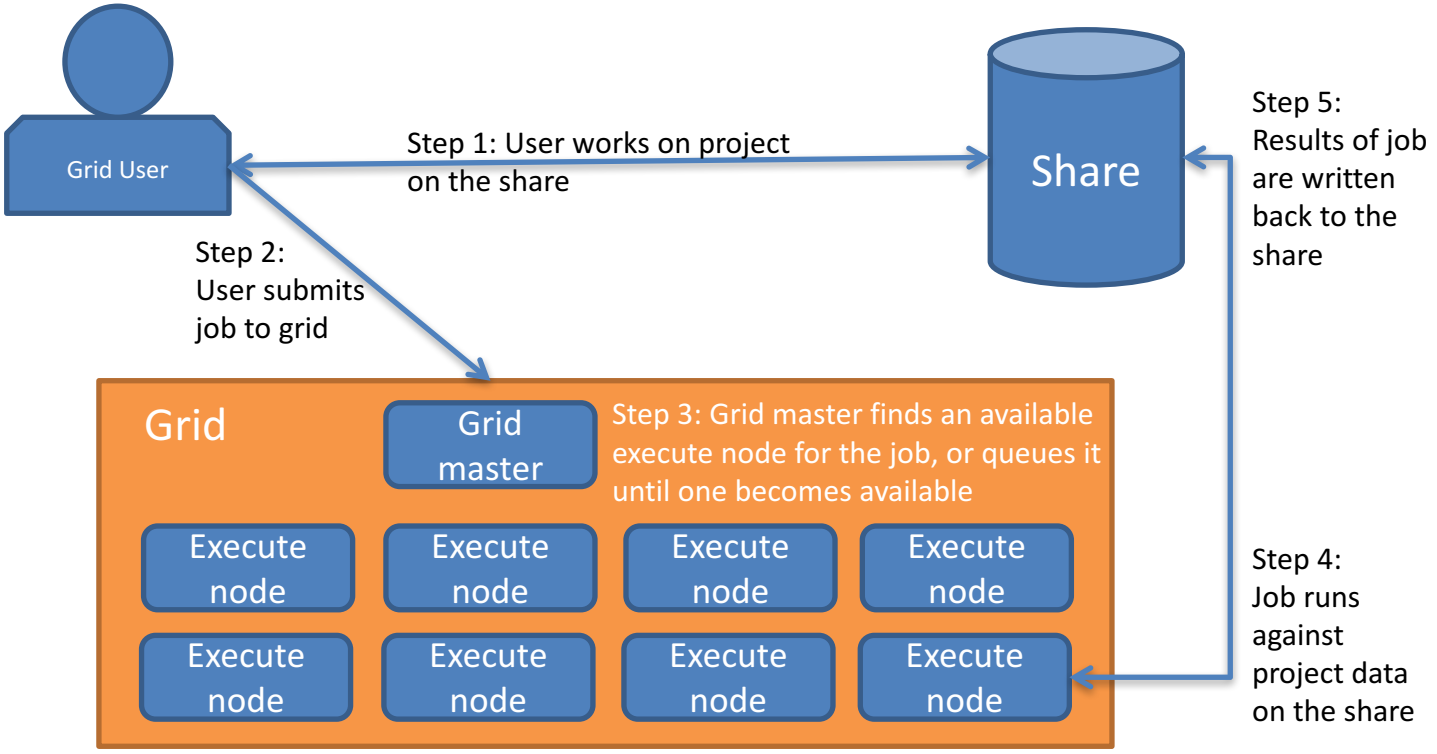
Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Grid Workflow



Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Detailed workflow



Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Why should I care?

- Grid computing is useful if you can break your analysis out into lots of small chunks which you can run in parallel

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Give me an example already!

- Project creates a model, in R, to run against a large dataset.
- The dataset splits out by customer contract, there are 38,000 of these.
- The model needs to be run 1000 times to reach the required level of confidence.
- 1000 runs on a single contract took about 1hour on a 32 core system with 60G RAM

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The rest of the example

- 38,000 hours of compute time – about 4.3 years on that single 32 core, 60G system.
- Because each of the 1000 runs is a discrete job, a grid allows us to scale that horizontally.
- So, using 1500 systems, the work would be complete is around a day.
- Cost for this type of instance from Amazon is $1.68ph =

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The downsides

- Traditional self hosted grids, in your own datacentre are rarely sized correctly.
- Don’t over engineer your job – remember the law of diminishing returns.

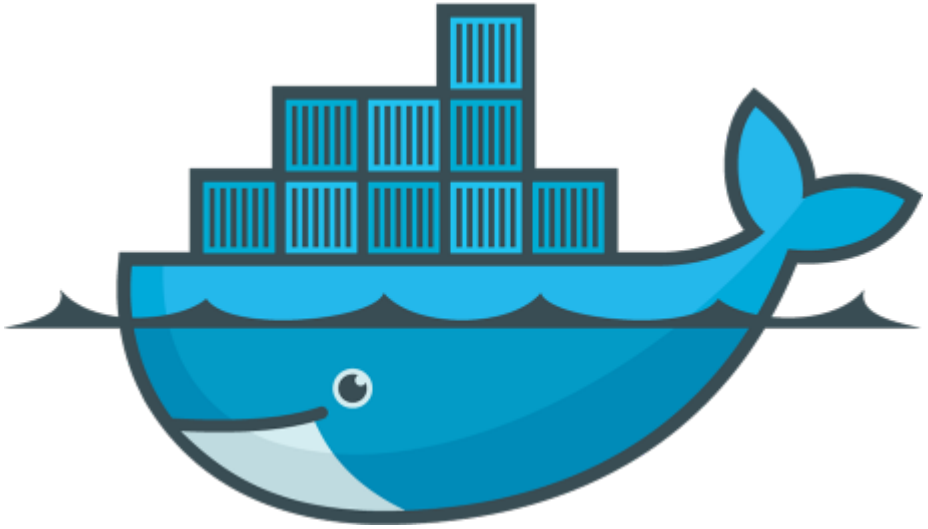
Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The future

- Expect to see more cloud based grids implementations.
- More hybrid grids will be used where some execution is internal but the cloud provides additional burst capacity
- I wouldn't be surprised to see someone offer grid execution as a service.

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Docker



Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

What is it?

- Docker uses Linux Containers to provide a lightweight wrapper for you applications.
- Provides image management and deployment services.
- Once your app is inside the container it will run anywhere docker runs.
- Everything your app requires to run exists inside the container.

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The docker project

- Is still young, but is maturing fast.
- dotCloud, the company that started the project has since rebranded as Docker Inc and made docker their primary focus.
- The project is receiving an unprecedented amount of attention and backing from across the industry
- Users include Facebook, Twitter, Netflix, Spotify etc

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Why should I care?

- Everything your app needs is in one place.
- Your project can be developed inside a container, the same container can be tested and exactly the same container can progress to production use.
- Need 5 instances of the same app? Just start the same container on 5 different servers

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Cut to the example

- Creating a remote execution environment
- A user creates and R script, script is pushed to the cloud, executed inside a docker container and the results are pulled back to the user.
- Docker provides resource isolation and the consistency of knowing that what's inside that container is fixed. Every time it starts it will be the same as when it was built.
- It's also easy to imagine having 3 containers with 3 different builds of R.
- You can take the container and run it anywhere, laptop, local server, cloud, etc.

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Another example

- We work on a lot of projects that use Rserve
- In order to more easily integrate R into generic web projects we use an Rserve HTTP API server which uses JSON
- Built inside a docker container the application suddenly becomes portable, repeatable and easily scalable.
- The same code is easily run on your laptop, on a local server, in the cloud.

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The downsides

- Not everything will run properly inside a container.
- Not everything is a good fit for containerisation.
- It's not yet well understood by most corporate IT teams
- It's a rapidly changing landscape.

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

The future

- We expect to see a lot more users to adopt this kind of containerisation approach
- ‘Web scale’ operations will continue to lead the charge in this area but the benefits will continue to trickle down.

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

Bringing it all together

- Cloud computing allows you to worry less about the infrastructure you use
- Using grid computing, run massively parallel batch style jobs
- Docker allows you to containerize your applications and can speed transition from dev through test and into production – you run exactly the same thing everywhere the container runs

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

QUESTIONS?

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com

THANKS!

Mark Sellors – Senior IT Consultant
msellors@mango-solutions.com