MATLAB Analysis of Pre-stack Seismic: Using MATLAB beyond the Geophysicist’s sandbox

Charles Jones* and James Selvage

7th October 2015
The following presentation contains forward-looking statements concerning BG Group plc's strategy, operations, financial performance or condition, outlook, growth opportunities or circumstances in the countries, sectors or markets in which BG Group plc operates, or the recommended cash and share offer by Royal Dutch Shell plc for BG Group plc announced on 8 April 2015. By their nature, forward-looking statements involve uncertainty because they depend on future circumstances, and relate to events, not all of which can be controlled or predicted. Although the Company believes that the expectations reflected in such forward-looking statements are reasonable, no assurance can be given that such expectations will prove to have been correct. Actual results could differ materially from the guidance given in this presentation for a number of reasons. For a detailed analysis of the factors that may affect our business, financial performance or results of operations, we urge you to look at the “Principal risks and uncertainties” included in the BG Group plc Annual Report & Accounts 2014. Nothing in this presentation should be construed as a profit forecast and no part of this presentation constitutes, or shall be taken to constitute, an invitation or inducement to invest in BG Group plc or any other entity, and must not be relied upon in any way in connection with any investment decision. BG Group plc undertakes no obligation to update any forward-looking statements.

No representation or warranty, express or implied, is or will be made in relation to the accuracy or completeness of the information in this presentation and no responsibility or liability is or will be accepted by BG Group plc or any of its respective subsidiaries, affiliates and associated companies (or by any of their respective officers, employees or agents) in relation to it.
Who we are

• Charles Jones
  – Processing geophysicist; converts raw data into bespoke data
  – Uses & writes algorithms in MATLAB
  – M.Sc. in exploration geophysics from the University of Leeds

• James Selvage
  – Geophysicist analysing data
  – Uses & writes algorithms in MATLAB
  – Demonstrated that MATLAB it is viable for solving large-scale geophysics problems
  – M.Sc. in exploration geophysics from the University of Leeds

• Programming is not our full time job
Our job is to evaluate the 3D subsurface

Images of remotely sensed subsurface
Talk Outline

• Challenge
  – Size and dimensions of seismic datasets
  – Timeframes of business

• Solution

• Benefits
What data do we use

• Acoustic Data called seismic – to image inside the earth

• Acquiring – Outsourced (3 months, $20 M, 100 km by 60 km)

• Pre-processing – Outsourced (6-12 months, $2 M, non-linear problem)

• Post processing (analysis) – Internal (3 months, $0.2 M)

• How do we use this data? – Decide where to drill ($50 M per well)
MATLAB history at BG Group

• Sandbox
  – Code from Academic Projects (SINBAD – UBC, SRB – Stanford, ETLP - HW)
  – Learn from academic papers and understand commercial algorithms

• Maturing
  – Replicated commercial analysis to improve efficiency on larger dataset
  – Prototype algorithm developed to improve on commercial offering
  – Decision point: [do nothing; another language; M codes data efficient]

• Production
  – Developed an improved approach to AVA analysis
  – High Level Language, maintainable and portable
  – Allow Non-professional programmers to develop their ideas
MATLAB history at BG Group

- **Sandbox**
  - Code from Academic Projects (SINBAD – UBC, SRB – Stanford, ETLP – HW)
  - Learn from academic papers and understand commercial algorithms

- **Maturing**
  - Replicated commercial analysis to improve efficiency on larger dataset
  - Prototype algorithm developed to improve on commercial offering
  - Decision point: [do nothing; another language; M codes data efficient]
MATLAB history at BG Group

• Sandbox
  – Code from Academic Projects (SINBAD)
  – learn from academic papers and understand commercial algorithms

• Maturing
  – Replicated commercial analysis to improve efficiency on larger dataset
  – Prototype algorithm developed to improve on commercial offering
  – Decision point: [do nothing; another language; M codes data efficient]

• Production
  – Developed an improved algorithm for analysis
  – High Level Language, maintainable and “putdownable”
  – Allow non-professional programmers to develop their ideas
Decision point – 2012

Make M codes production ready

Multi-terabyte datasets

Embarrassingly parallel

Parallel processing

Maintainable code

MATLAB solution

MATLAB MDCS or MATLAB Compiler

MATLAB MDCS

High Level Programming Language
Decision point – 2012

Make M codes production ready

- Multi-terabyte datasets
- Embarrassingly parallel
- Parallel processing
- Maintainable code

MATLAB solution

- MATLAB MDCS or MATLAB Compiler
- MATLAB MDCS
- High Level Programming Language

What are these multi-terabyte datasets?
Acquisition (3 months, $20 M)

Streamer [12] (receivers)

Airguns (source)

PGS Ramform Titan in Tanzania
Acquisition (3 months, $20 M)

- 280,000 shots
- 20 - 30 TB

Processed to form a 3D cube of data that "images" the subsurface
Pre-Processing (6 – 12 months, $2 M)
Analysis – (3 months, $0.2 M)

Seismic has limited vertical resolution

70m

3 dimensions
Analysis – (3 months, $0.2 M)

3 dimensions
Analysis – (3 months, $0.2 M)

increasing incidence angle

4 dimensions X, Y, Z, Angle

This is the multi terabyte dataset; upto 40 TB
These datasets are getting larger ...
Outline

• Challenge
  – Size and dimensions of seismic datasets
  – Timeframes of business

• Solution
  – Read and manipulate large binary datasets into MATLAB [typecast, vectorised array operations, structures]
  – Embarrassingly parallel processing [MATLAB Compiler + Manual coding]
  – Maintainable code [High Level Programming Language]

• Benefits
Solution for large datasets

Scan SEG-Y

SEG-Y Read

Algorithm

SEG-Y Write

Dataset

SLURM Cluster Manager

Compile Function

SLURM submit
Index dataset: Scan segy

- Outputs a meta file for the input data
- Structures for meta data storage
- Compressed binary index file
Index dataset: Scan segy

For a 310 Gbyte segy file:

Meta data – mat file – 2 Kbytes

Compressed binary index file – 842 Kbytes (9.3 Gbytes)

Linear IO Read in 10 Mbyte chunks @ ~350 Mbytes / second
Library to make data parallel codes

SEG-Y Read

Algorithm

SEG-Y Write

\[
\begin{align*}
\text{seg_read_binary} & \quad \text{seg_index_byte_finder} \\
\text{node_segy_read} & \quad \text{node_segy_write} \\
\text{int_grad_inv_proj} & \quad \text{load_wavelet_spatial} \\
\text{low_amp_mute} & \quad \text{bandpass_filter} \\
\text{water_bottom_picker} & \\
\text{gather_compress_ilxl_bytes} & \\
\text{trace_compress_ilxl_bytes} & \\
\end{align*}
\]
Parallel processing

Seismic dataset divided into blocks.

• Submit compiled exes to cluster using SLURM
  – Jobs can co-exists with non-MATLAB exe
Parallel processing

>> node_slurm_submit2013(algorithm_name, job_meta_path, slurm_part, n_cores, varargin)
Parallel processing (10 TB, 0.5 days, 576 cores)
Parallel processing (10 TB, 0.8 days, 576 cores)
Parallel processing (10 TB, 1.2 days, 576 cores)
Parallel processing  (10 TB, 1.8 days, 576 cores)
Parallel processing (10 TB, 2.1 days, 576 cores)
What was the time benefit?

• Acquiring – 3 months, $20 M

• Pre-processing – 6-12 months, $2 M

✓ Post processing – 3 months, $0.2 M, now 3 days
Outline

• Challenge
  – Size and dimensions of seismic datasets
  – Timeframes of business

• Solution
  – Read and manipulate large binary datasets into MATLAB [typecast, vectorised array operations, structures]
  – Parallel processing [MATLAB Compiler]
  – Maintainable code [High Level Programming Language]

• Benefits
  ✓ Run Time [3 months to 3 days]
  ✓ Very quick to adapt as datasets are different every time
  – Further developments
Led to more developments

Digi (Dynamic Int & Grad Inversion)

Anom. Spotter & Anom. Connector

Mica (Ratno)

Classification (Erik)

Water column (Dan)

Dip Tensors (James)

sandbox production expansion
Next steps – more ambitious – 2014 and beyond

- Acquiring – 3 months, $20 M
- Pre-processing – 6-12 months, $2 M

✓ Post processing – 3 months, $0.2 M, now 3 days
International Inversion Initiative

• SINBAD research group at University of British Columbia (Canada)
  www.slim.eos.ubc.ca

• Fullwave consortium at Imperial College London (UK)

• Senai Cimatec in Salvador (Brazil)

• Solving the non linear inverse problem on the raw field data rather than pre-processed data
International Inversion Initiative

- Funded by BG Group and SENAI/Fieb
- MATLAB distributed computing server – 4000 workers
- 17200 cores at Senai Cimatec
- 132 TB of memory
- No. 165 on July 2015 Top500
SINBAD consortium MATLAB codes

- Object-oriented-programming and parallel-programming libraries that implement linear algebra operators
- Libraries are organized into topical toolboxes, e.g. FD frequency modelling, compressive sensing
Matrix-free framework

1) SPOT linear-operator toolbox brings MATLAB’s built-in matrix notation to problems for which explicit matrices are not practical.

http://www.cs.ubc.ca/labs/scl/spot/

\[
n = 1000; \quad x = (1:n)'; \\
F = \text{opDFT}(n); \quad \% \text{create a DFT operator} \\
s = \text{sqrt}(n) * F * x; \quad \% \text{eigenvalues of } C \\
C = \text{real}(F' * \text{opDiag}(s) * F); \quad \% C \text{ is } 1000 \times 1000 \text{ 34 Kbytes not 7812 Kbytes} \\
z = C' * y; \quad \% \text{apply the adjoint of } C \text{ to a vector}
\]

2) pSPOT extends the SPOT operators to act on MATLAB's distributed vectors using MDCS
SINBAD parallelism

• Data-space parallelism,
  – Distributed arrays, each worker processes part of either in-core distributed array or out-of-core data. (MDCS or Compiler)

• Model-space parallelism,
  – Model-domain decomposition, distributed arrays and MPI inter worker communication. (MDCS)

• Both use MATLAB's intrinsic or Mex multi-threading
SINBAD problem sizes

• Full waveform inversion
  – Frequency domain finite difference wave propagation
    – 9.3 TBytes, 17200 cores = 84 days
  – Time Domain FD
    – 23 TBytes, 17200 cores = 20 days

• Least squares Reverse Time Migration
  – 20 TBytes, 17200 cores = 42 days

• Simultaneous source acquisition
  – 2 Tbytes, 17200 cores = 50 days
Next challenge

- Resilience – heterogeneous MPI
- Parallel data distribution
Conclusion

• MATLAB is being used extensively by BG Group

• Started as a sandbox environment, now used for routine production

• Collaborating on more ambitious problems
Acknowledgements

• BG Group for permission to publish this work

• BG Co-coders – Ratnadwip Ghosh, Dan Bright, Jon Edgar, Simon Wrigley

• BG Business sponsors – Ian Turner, Hamish Macintyre

• University British Columbia – Felix Herrmann, Tim Lin, Henryk Modzelewski

• Imperial College London – Mike Warner, Gerard Gorman
Open Performance portable Seismic Imaging

- Intel® Parallel Computing Centre at Imperial College London and SENAI CIMATEC
- Industrial partners – Intel and BG Group
- Implements FD and FE wave propagation in high level language
- Automatically optimises for different architectures
- Already outperforms industrial code on Xeon and Phi platforms
- [https://github.com/opesci](https://github.com/opesci) - opensource