A L D R I X G R O U P

Open Source Systems Sdn. Bhd.(MSC - Status)

White Paper

HIGH PERFORMANCE COMPUTING

ALDRIX SOLUTIONS

GRID COMPUTING

BIOINFORMATICS

NUMERICAL WEATHER PREDICTION

VIRTUAL REALITY

June 2005

Block D2 UPM-MTDC, Lebuh Silikon, 43400 Serdang, Malaysia • tel: +603 8656 0139 • fax: +603 8656 0132
www.aldrix.net

Table of Contents

Front Cover: The Belum National Forest, Perak, Malaysia

Introduction

The Aldrix Group, through its company, Open Source Systems Sdn. Bhd. (OSS), has been in the business of providing distributed computing solutions in Malaysia since 1998. Our focus has always concentrated on providing leading-edge technologies that are reliable and reasonably priced. We delivered the first commercial Beowulf Linux cluster in Malaysia in 1998 and our core solutions are based on a wide selection of Linux/Open Source tools. We pride ourselves in being vendor-neutral, preferring to concentrate on the requirements of our clients and providing customized solutions.

In this white paper, we aim to share our experiences in using the technologies currently available for the High Performance Computing and Grid Computing industry. Over the years, as our solutions progressed with the maturity of the various technologies available, we have acquired a certain level of skillsets in the configuration and customization of systems which are tailor-made for our clients, which mainly consists of universities, research agencies and laboratories.

Among the topics that we will be cover are:

- HPC Technological Overview
- Leading-edge technologies applied in our High Performance Computing solutions
- Our Core Competency and Expertise
- Case Studies: The Success Stories

HPC Technological Overview

Any informatics facility requires proper planning to ensure that the facility can meet the demands of current as well as future needs of their users. This section will focus on the current as well as possible future trends in several areas that are important in research computing and usage of HPC in the commercial world.

Software

The first and most important component to consider is software. Some applications were written to run only on specific hardware platforms but most many applications today can run on multiple platforms.

The availability of open source software in some research environments for example bioinformatics and computational chemistry has leveled the playing field for researchers in that arena due to its low acquisition cost. In addition, open source applications are more easily recompiled to run on multiple hardware platforms, reducing the need to purchase specific systems for specific software.

The idea of open source is congruous with research in general. A researcher may write an application and make it available to other researchers in his or her field. Most people will use the algorithm but a handful will improve it and pass back the improvements to the community. Over time, the application will flourish and become a known standard.

There are many examples of open source applications that started in the academic world. Some of the more popular ones include Apache, Samba, BLAST, FastA, EMBOSS and the various Linux and BSD distributions. However, many applications that have ended up in the open source community were started by commercial companies as well. These include the Sun N1 Grid Engine (SGE), OpenOffice, Mozilla, Eclipse and others.

In addition to open source, there are however many important commercial software that have been ported to run on the open source platform. Databases (Oracle, Sybase, DB2), computer aided design/engineering applications (Pro/Engineer, Unigraphics, Nastran), IC/chip design (CADENCE), mathematical modeling (Matlab, Mathematica) and other commercial software have been ported to run on the Linux platform. Many of these applications are also “cluster ready” i.e. they can take advantage of an HPC cluster with no changes.

Closely related to open source software especially are the compilers used to compile these software. The vast majority of software do not require a lot of performance or optimization, however for those that do, the choice of compilers will greatly determine the level of performance that a researcher can squeeze out of the available hardware.

Any research institution can do respectable research by purely using open source software but this requires the scientists in the institution to be thoroughly familiar with the open source environment. In the real world, we believe in using the best tool for the job. This could mean open source for some areas and commercial software for other areas; however, the usage of open source software however is expected to increase worldwide.

High Performance Computing (HPC) Clusters

HPC Clusters have been growing in prominence over the past decade due to the advances in personal computer technology. Seven out of the top ten supercomputers in the Top 500 Supercomputers list are cluster type systems. An HPC cluster is a collection of two or more individual machines connected together over a network in such way that it behaves as a single computer.

Illustration 1: An early Beowulf cluster at NASA (1995)

HPC Clusters literally have three major components, the hardware and CPU architecture, the operating environment/system and the interconnect that connects the multiple nodes with each other.

HPC Cluster Hardware

On the hardware side of HPC Clusters, the current trend today seems to be the x86 based systems. This literally means Intel or AMD based systems. These systems provide good price/performance ratios and can scale reasonably well. There are numerous companies providing x86 based HPC Clusters from large players such as Dell, IBM, Sun and HP to smaller, specialized outfits such as Appro, Racksavers, Linux Networx and Penguin Computing.

Non-x86 based HPC Clusters do exist, however they are less popular and focus mainly on applications that are either optimized for an alternate platform or do not exist on x86 platforms. These clusters can be based on Intel Itanium, Sun UltraSparc, HP Alpha or IBM PowerPC based.

32-bit IA32 and 64-bit x86-64 architectures

The IA32 and x86-64 architectures are also popularly known as the x86 architecture. For the foreseeable future, the x86 architecture will continue to be the choice of many HPC Clusters. The new kid on the block, however is the 64-bit x86 architecture, pioneered by AMD in 2003 and followed closely by Intel a year later. These systems continue to provide excellent price/performance ratio but have the advantage of 64-bit addressing, allowing them to address significantly larger than the old 4GB limit. Advances in hardware design on the AMD Opteron platform has also provided a significant boost in memory access speeds, improving the overall performance of these systems even higher.

Illustration 2: Albert, the 530-CPU AMD Opteron based HPC used for CFD by Sauber-Petronas (Switzerland, 2004)

PowerPC architecture

In addition to x86-64 systems, the 64-bit IBM PowerPC G5 CPUs, in the form of the Apple Power Macintosh G5 is beginning to make headway into the HPC cluster circles. The PowerPC G5 is the cheaper, lower-end version of the IBM POWER4+ CPU providing excellent performance for the price and better floating point performance than most x86 systems. The first 64-bit PowerPC G5 based HPC cluster at Virginia Institute of Technology went straight to the top of the HPC cluster heap of the November 2003 Top 500 list, ranking in at number 3 (numbers 1 and 2 are customized supercomputer style systems), beating the Intel Xeon, Intel Itanium and AMD Opteron based systems (which interestingly enough were numbers 4, 5 and 6 respectively in the November 2003 Top 500). The PowerPC G5 cluster also posted the best price/performance ratio in the list and was second only to the Itanium for the best performance/CPU ratio. While IBM provides PowerPC G5 and POWER5 based systems in their range of products, the future of the Apple Power Macintosh G5 and XServe G5 systems are unclear as

Apple has announced that they will be moving away from the PowerPC processors beginning in 2006.

IA64 architecture

The fast but largely unpopular Intel Itanium IA64 platform is a worthy platform for any HPC system. Unfortunately, its somewhat higher price and perceived lack of compatibility with existing 32-bit software has prompted people to go towards cheaper platforms including x86-64 and IA32. AMD and Intel's support for the x86-64 platform however has provided many people with 64-bit capabilities and the IA64 platform unfortunately continues to be a distant third or fourth choice for many HPC cluster users.

HPC Interconnects

Interconnect hardware is the communication hardware used to connect the different nodes of an HPC cluster. These can range from commodity ethernet hardware (fast or gigabit ethernet) to specialized cluster interconnects such as Infiniband or Quadrics to custom high-speed NUMAflex connections used on SGI Origin and SGI Altix systems.

Ethernet

Fast/gigabit ethernet is a popular option for many small to medium sized clusters. These systems generally run embarrassingly parallel jobs that require little or no data from a central server, e.g. in a rendering farm. The price is cheap and the ethernet hardware is generally built in into the system board making installation and configuration painless. Ethernet provides bandwidth of between 100Mbit/second to 1000Mbit/second with latencies in the 10-100 millisecond range.

However, while fast ethernet or gigabit ethernet is sufficient for a small cluster of say up to 16 nodes, anything beyond that will benefit tremendously from faster interconnects. To paraphrase a local researcher, Dr. Rosni Abdullah of Universiti Sains Malaysia, Penang, slower interconnects will cause the nodes to literally spend more time talking than doing work. Even smaller clusters running communications intensive applications for example molecular dynamics simulations or weather modelling applications will benefit from faster interconnects.

Myrinet

The market leader in specialized cluster interconnects today is Myrinet. Myrinet provides bandwidth of more than 2000Mbit/second with latencies around 10 microseconds. This is double the bandwidth of gigabit with 10-100 times better latencies. Myrinet works with most hardware and software platforms out there including IA32, x86-64, Sparc and PowerPC.

Quadrics

The emerging contender to Myrinet is Quadrics. Originally developed to connect the Compaq Alphaserver clusters, the Quadrics interconnect provide better bandwidth than Myrinet at around 4000Mbit/second and similar latencies. The Quadrics software includes a shared memory library; allowing processes to share memory on different nodes across the Quadrics inter-

connect. Quadrics is supported on many hardware platforms as well, including x86 and PowerPC.

Infiniband

Over the past several years, Intel has hyped the Infiniband interconnect as the next generation interconnect to replace all other interconnects. While this has not happened yet, Infiniband has found a niche market in the HPC arena where it provides more than 5000Mbit/second bandwidth and sub 10 microsecond latencies. Being an open standard, Infiniband is fast becoming the "commodity" standard of high-speed interconnects with multiple vendors providing hardware at lower cost.

Our Core Competency and Expertise

Automated Linux Cluster Installation

(NPACI Rocks Linux Cluster Distribution)

The main operating system that we are using for our Linux clusters is the NPACI Rocks Linux Cluster Distribution (http://www.rocksclusters.org), which is developed by the San Diego Supercomputing Center (SDSC). The Rocks Cluster Distribution is currently being widely used as the operating systems for Linux clusters across the globe as it offers great features such as easy installation of clusters, the tightly coupled packaging of various clustering software and tools in the installation of nodes and specialized tools, created specifically for cluster management.

OSS' in-house HPC developers have been participating actively in the Rocks mailing list and the tweaking of the system to suit the requirements of various clients. This familiarity has allowed OSS to acquire a competitive edge in the local HPC market as we are able to provide solid and satisfactory HPC system to all our customers, built on the high performance and reliability of the Rocks Linux Cluster Distribution.

Software Development for Parallel Computing

The common parallel programming libraries used for the development of HPC software are Message Passing Interface (MPI) and Parallel Virtual Machine (PVM). As explained in the paper entitled "Goals Guiding Design: PVM and MPI " presented by William Gropp and Ewing Lusk of the famed Argonne National Laboratory, PVM and MPI are essentially designed to solve different problems. Hence, there is a need to be familiar with both sets of tools as different requirements will involve different solutions. However, the implementation of MPI applications seem to be more widespread currently in HPC systems. This is due to the fact that PVM is the effort of a single research group (The Oak Ridge National Laboratory's Computer Science and Mathematics Division) whereas MPI is implemented by the MPI Forum (http://www.mpi-forum.org/), a diverse group of implementors, programmers and end users.

We provide software development for parallel computing needs to our client. With our knowledge on parallel computation, we provide consultation to our clients in parallelizing their applications where possible. Currently, our focus is on scientific applications as the computation algorithms for fields such chemistry, physics and biology can be parallelized to produce results in shorter time using MPI and HPC systems.

Grid Computing Technologies

Grid computing is essentially the management of computing resources, that are most of the time geographically dispersed, in order to achieve greater efficiency in the utilization of these resources. As almost all of these systems are interlinked together via high-speed networks, new middleware are being created to take advantage of these features. Among them are grid-enabling tools such as the Globus Toolkit and job schedulers such as the N1 Grid Engine (formerly known as Sun Grid Engine).

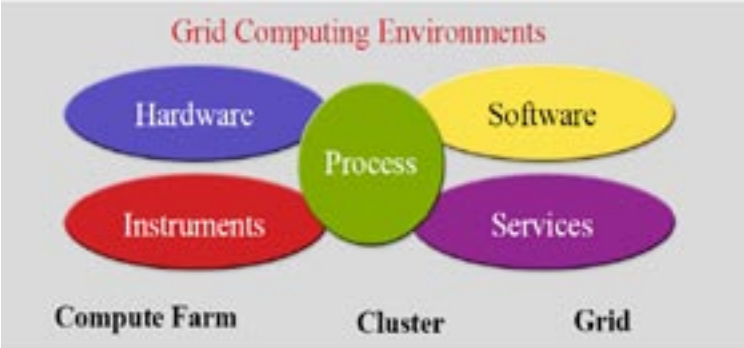

Illustration 3: The Main Components of a Grid (Enterprise Computing With Sun Grid Engine. Charu Chaubal, Sun Microsystems. 2004 ClusterWorld Conference and Expo)

The N1 Grid Engine (http://gridengine.sunsource.net/) is an open source job scheduling toolkit that was developed by Sun Microsystems. It allows users to submit their processing jobs to the grid. These jobs form a queue in the Grid Engine and the engine will then send these jobs to the most suitable machine in the grid to do the processing on a first-in-first-out basis (default behavior). The machine selected to do the job is based on the grid engine's algorithm (taking into account the various machines' load levels etc).

The Globus Toolkit (http://www.globus.org) is a set of software, libraries and APIs developed by the Argonne National Laboratory that is used to gridify (grid-enable) applications to run on a grid, letting users share computing power, data and other tools securely.

With the combination of Globus and N1 Grid Engine, we can easily extend the grid implementation from the departmental grid level to enterprise and global. We have worked with multiple parties including Sun Microsystems and Microsoft Corp. to implement grid computing technologies. We were even invited to showcase our work in implementing HPC clusters using Microsoft Windows 2003 x64, Condor, MPI and Globus at Microsoft's Innovation Day 2004.

However, the research in grid computing is an on-going process and there are many exciting technologies and concepts to be discovered, tried and tested. Therefore, the Aldrix Group is very proud to be part of this exciting period in grid research.

Illustration 4: Levels of Grid Deployment
(Grid Computing with SGE/SGEE and Globus. Dr Simon See, Sun Microsystems. INDO-UK Workshop on e-Science, 2004)

Optimization, 64-bit Computing and Benchmarking

We have always believed in using the latest technologies in our solutions. Therefore, we have been experimenting with the AMD64 platform (using the AMD Opteron processors) in our HPC solutions since the launch of the Opteron processor in 2003. We believe that this early adaptation to the first 64-bit x86 computing platform has given us the advantage over other HPC implementors in Malaysia as the Opteron processors performance and our ability to harness that performance has increased our clients' level of confidence in our capabilities. We have had experience on 64-bit Linux on the Sun Sparc and IBM POWER line of machines before and the migration to 64-bit Linux on the AMD Opteron proved to be very easy.

Time and again, we have successfully run many applications on the x86-64 platform using AMD Opteron technology. For example, we were among the first to have successfully run the Numerical Weather Prediction (NWP) software, MM5 (Mesoscale Model) on a 64-bit Linux cluster.

As Linux clusters gain popularity in the High Performance Computing arena, many commercial software makers are developing their applications. One important area of software in Linux clusters are the compilers. Compilers are the applications that convert the source codes into binaries and libraries that will carry out the instructions that the developers has created. Although normally hidden from the end-user, the role of the compiler is very important in the never-ending chase for greater performance. Therefore, our experiences in using the various compilers in Linux, both on the commercial side with PathScale, the Portland Group and Intel or on open source with the GNU compilers, have helped our clients to find that elusive increase in performance in their applications.

We have always believed that in the field of High Performance Computing, action does speak much better than words. Therefore, we have always used a more proactive role to convince our clients of our capabilities and solutions. The easiest way to do so is to provide our clients with concrete benchmarking information, more convincing when the applications that we benchmark are relevant to the client's area of interest. Therefore, we have developed in-house expertise in the benchmarking the scientific applications that our clients use such as High Performance Linpack (HPL) for general benchmarking, Gromacs and Amber (molecular dynamics in computational chemistry), MM5

(Mesoscale Model) for Numerical Weather Prediction (NWP) and numerous bioinformatics applications.

Case Studies: The Success Stories

The following are four unique clusters that the Aldrix Group has installed. Excluding some private clusters in the oil and gas industry, we believe that these clusters are also the biggest in Malaysia thus far. Therefore, the Aldrix Group still holds the enviable record of installing the top three biggest research clusters in Malaysia.

Malaysia Genome Institute (MGI)

In 2000 when Linux, let alone Linux clusters, was barely known in Malaysia, we were already in discussions to deliver a 24-processor HPC cluster based on the Intel Pentium III Xeon to the Malaysia Genome Institute (MGI). Known as the Centre for Gene Analysis and Technology (CGAT) then, MGI is still using the cluster extensively to run their sequence analysis, gene prediction, basecalling/assembling and other bioinformatics applications (such as BLAST, HMMER, STADEN, EMBOSS, GROMACS etc.) for genomics and other related research. Our locally developed Open BioLIMS (Laboratory Information Management System) application feeds the cluster with data directly as well.

Currently, the Aldrix Group still provides maintenance and administrative services to MGI for this cluster. After almost four years of existence and several rounds of hardware upgrades, we are currently working closely with MGI to significantly increase the performance and capacity of the cluster from their current 60+ processors, via new hardware installations that will ensure that the researchers will be able to get their results even faster. The cluster has recently been strengthened with a 5-terabyte storage array and multiple new AMD Opteron based compute nodes. Challenges that we have faced or are facing include the management of storage and backups for MGI's high throughput DNA sequencing facility which uses the cluster, integration of multiple hardware platforms (IA32, IA64, Sun Sparc and SGI MIPs) to provide seamless access for the researchers and integrating Globus and grid computing into their infrastructure for collaborative work with other local universities.

The knowledge and experiences gained in this installation set the foundation and impetus for greater things to come as we continue to create new records in installing bigger and better clusters nationwide.

Malaysian Institute of Microelectronics Systems Berhad (MIMOS)

At the end of 2004, we successfully won a tender to supply the Malaysian Institute of Microelectronics Systems, MIMOS, the nation's most powerful research supercomputing facility thus far. It is 68-processor Opteron-based system with 136GB memory and a multi-terabyte storage array. We are proud of this solution as it is entirely configured and customized by the Aldrix technical team based on the Shanghai Supercomputer Center and other systems worldwide.

We faced different challenges for this system, many related to security, accounting and policy issues as MIMOS provides free access to this HPC system for Malaysian researchers. This cluster represents another milestone for us, as a company and for the nation of Malaysia as well. It heralds the beginning of exciting activities in the High Performance Computing and Grid Computing arena, something that we have been promoting and pushing very hard for the past few years in the local research circle.

MIMOS has also been given the mandate by the Ministry of Science, Technology and Innovation (MOSTI) or Malaysia to set up a facility to provide grid computing research and facilities in the Malaysian academic arena and intends to expand their HPC system to 512-processors to provide a strong national supercomputing center. Therefore, we are also providing technical consultation to MIMOS to help formulate the roadmap that will implement the National Grid Computing Facility.

Malaysian Meteorological Services (MMS)

In line with our sincere interest in aiding national efforts, Aldrix Group is currently building a High Performance Computing cluster for the Malaysian Meteorological Services or also known as Jabatan Perkhidmatan Kajicuaca Malaysia, JPKM. The Malaysian Meteorological Services, under the administrative control of the Ministry of Science, Technology and Innovation (MOSTI), is responsible for tasks related to climate, weather conditions and seismic activities in Malaysia and the region.

Numerical weather prediction is one of the most widely used applications in the HPC field today. Weather prediction in the equatorial region can be particularly complex and new techniques and algorithms are implemented continuously to provide better forecasts.

The Aldrix Group, through Open Source Systems, won the opportunity to provide and install the cluster hardware, as well as employ our experience and technical skill in customizing the setup. Customization work includes optimizing the Operating System, compilers, and applications, as well as porting the required software components to run seamlessly and effectively on the system.

The challenges for this project is different as this will be a production system with continuous data being fed from weather stations worldwide with calculations running 24x7. Data archival is absolutely important for this environment as well. Future enhancements will include new data types from satellite, doppler radar, etc., research into the WRF models and possible implementation of ensemble forecasting.

Universiti Industri Selangor (UNISEL) *i*VR Lab

This system is a different type of HPC cluster, designed to power a graphically intensive virtual reality (VR) system for engineering and virtual prototyping at the Engineering Faculty, UNISEL. Installed as the back-end server for a two-channel VR system, each node of the 4-node cluster powers one projector to provide a distributed real-time OpenGL based graphical rendering system. With a combined video and texture memory of 512MB, the system provides the same level of performance of traditional VR systems from SGI and HP but at a fraction of the cost.

This cluster's unique ability comes from the combination of hardware and software that was conceived originally at Fraunhofer Institute IAO and currently being developed and marketed by ICIDO GmbH, Stuttgart. OSS has been an exclusive partner with Fraunhofer IAO/ICIDO since year 2000 and have been fully trained to manage, implement and support the solution.

The ability to view and display other 3D models in the field of bioinformatics and architecture is an additional bonus. Used in major automotive and engineering companies worldwide including Daimler-Chrysler, BMW, Ford Motor Co., Volvo and Airbus, this Linux-powered VR solution allows researchers in UNISEL to be at the forefront of engineering and visualization technology.

About The Aldrix Group

The Aldrix Group was founded in 1998 by its CEO, Dr Azman Firdaus Shafii. It consists of two main companies, Open Source Systems Sdn Bhd and Intelligent Urbans Systems Sdn. Bhd (IUS). OSS is the Malaysian pioneer in Linux and Open Source software. The company developed Malaysia's first Linux-based supercomputer and was awarded the MSC-status for this effort. OSS is also a local pioneer in Bioinformatics and Virtual Reality, having established Malaysia's first Bioinformatics Linux cluster at the National Genomics Laboratory in Universiti Kebangsaan Malaysia, a Linux-driven VR CAVE facility at Universiti Malaysia Sarawak and an iVR POWERWALL system at UNISEL. For several years, OSS, through Dr Azman, has been an active contributor to various national biotechnology task forces and work groups mapping the biotechnology agenda for Malaysia. IUS has developed Perbadanan Putrajaya's City Control Centre that manages intelligent traffic transportation, public information and emergency responses as well as public facilities including the Putrajaya web portal (www.putrajaya.net.my). This test-bed integrated project is believed to be the first of its kind for any Local Authority in the world.

As a Colombo Plan Scholar, Dr Azman studied at the University of Sydney, Australia where he graduated with the ICI Australia and New Zealand Prize for Chemical Engineering. He then went to Aston University, UK, to do his PhD in Computer Process Control. After a five-year stint at Universiti Malaya, Dr Azman served Golden Hope Plantations Berhad for ten years. During this period, he initiated and led many pioneering technology advancements in the processing of cocoa, natural rubber, latex, tropical fruits, palm oil, rubberwood products and downstream palm oleochemicals. He was also instrumental in establishing Golden Hope's first palm oil refineries in Vietnam and China. In 1991-1993, he took leave to do his MBA at Harvard Business School, USA. His last position in Golden Hope was Group Director, Manufacturing.

Dr Azman left Golden Hope to manage the Penang Bridge Privatisation project, with eventual listing into Intria Berhad, and managing toll operations and property development projects under Metacorp Berhad. Under Dr Azman's leadership, Intria acquired a world-class British engineering and construction firm, Costain PLC; and subsequently he became the first non-British Chairman of this company in 1996, listed on the London Stock Exchange.

In 1998, Dr Azman decided to pursue his technopreneurial vision. He founded two ICT companies, an R&D driven company, Open Source Systems Sdn Bhd and Intelligent Urban Systems Sdn Bhd. He also served as Adjunct Professor of Strategic Management in the Malaysian Graduate School of Management, Universiti Putra Malaysia in 1999 – 2001. Currently, he is an Academic Advisor to Universiti Industri Selangor (UNISEL) as well as an ICT and Biotechnology Panel Advisor for the State of Perak.

References

I. Foster, C. Kesselman, S. Tuecke. 'The Anatomy of the Grid: Enabling Scalable Virtual Organizations', International J. Supercomputer Applications, 15(3), 2001.

C. Chaubal, Sun Microsystems, 'Enterprise Computing With Sun Grid Engine', ClusterWorld Conference and Expo, 2004

S. See, Sun Microsystems, 'Grid Computing with SGE/SGEE and Globus', INDO-UK Workshop on e-Science, 2004

Parallel Virtual Machine (PVM). http://www.csm.ornl.gov/pvm/

Message Passing Interface (MPI). http://www.mpi-forum.org/

MPICH. http://www-unix.mcs.anl.gov/mpi/mpich/index.htm

LAM/MPI. http://www.lam-mpi.org

Globus Toolkit and Grid Computing the Globus Way. http://www.globus.org/

Malaysia Genome Institute (MGI). http://cgat.ukm.edu.my

Malaysian Institute of Microelectronics Systems(MIMOS Berhad). http://www.mimos.my

Malaysian Meteorological Service (MMS). http://www.kjc.gov.my

Universiti Industri Selangor (UNISEL). http://www.unisel.edu.my

Contact Persons

Dr. Azman Firdaus Shafii, Founder Chairman and CEO dafs@aldrix.net

Farul Mohd. Ghazali, Manager, Systems & Bioinformatics farul@aldrix.net

Nicholas Low, Manager, Business Development nicholas@aldrix.net