Redbooks Paper

Yinhe Cheng
Joyce Mak
Chakarat Skawratananond
Tzy-Hwa Kathy Tzeng

Scalability Comparison of Bioinformatics for Applications on AIX and Linux on IBM eServer pSeries 690

Abstract

The scalability performance of several critical bioinformatics applications, including BLAST, FASTA, Smith-Waterman, HMMER, ICED and Mascot, was evaluated on AIX® 5.1 and SUSE Linux Enterprise Server 8 for IBM® eServer pSeries®. These applications are either threaded or OpenMP-enabled parallel implementations. The comparison was conducted on the same LPAR of a p690 with 16 CPUs. The results suggested that the scalability of these applications on AIX and Linux are comparable, in general.

Introduction

As Linux gained popularity, Linux on pSeries became a key element of IBM eServer Linux. By providing Linux on pSeries, Linux users running on Intel® architecture machines have the option to move up to more powerful pSeries systems.

However, it is widely believed that Linux does not currently scale to efficiently handle large SMP (1). It was our intention in this study to evaluate how well the leading bioinformatics applications scale on the current offering of Linux on pSeries 690.

Four bioinformatics applications (BLAST, FASTA, Smith-Waterman and HMMER), one Microarray analysis application (ICED), and one proteomic application (Mascot) were analyzed on AIX 5.1 and SUSE Linux Enterprise Server V8 for p690. IBM VisualAge® compilers were used, and the same compilation flags were applied for both AIX and Linux platforms.

System configuration

The system configuration used in this study was a logical partition (LPAR) on a 1.1 GHz POWER4™ p690. LPARs are user-defined system divisions consisting of CPUs, memory, and I/O slots that are a subset of the available resources within a system, regardless of their locality. Each LPAR runs a specific instance of an operating system, the same way it runs on

standalone servers. Logical partitioning is very transparent to user applications. The LPAR created for this study had 16 processors and 32 GB memory.

The following software configurations were used.

- Linux

  SUSE Linux Enterprise Server 8.0

  - kernel 2.4.19-108
  - IBM VisualAge C++ v6.0.0.0

- AIX

  AIX 5L™ v5.1 ML-4

  - 32-bit kernel mode and zero large page
  - IBM VisualAge C++ v6.0.0.3

Benchmarks and results

Benchmarks of the following applications were measured in our study:

- BLAST - thread implementation
- FASTA - thread implementation
- Smith-Waterman - thread implementation
- HMMER - thread implementation
- ICED - OpenMP implementation
- Mascot - thread implementation

In the following sections, the performance and scalability comparison of these six applications on both AIX5.1 and on SUSE Linux are presented on the same hardware.

BLAST

BLAST (Basic Local Alignment Search Tool) is a set of widely used tools for rapid sequence similarity search against both nucleotide and protein databases (2). Several variants of BLAST can be distinguished by the type of the query and database (see Table 1). In this study, blastn, blastp, and blastx from BLAST 2.2.6 were used for evaluation. For each program, two searches of different lengths were conducted.

Table 1 BLAST programs

Program	Query	Database
blastp	Protein	Protein
blastn	Nucleotide	Nucleotide
blastx	Nucleotide (translated to protein on the fly)	Protein
tblastn	Protein	Nucleotide (translated to protein on the fly)
tblastx	Nucleotide (translated to protein on the fly)	Nucleotide (translated to protein on the fly)

Blastp

In this study, human protein sequences were used to search against an nr database (1508490 sequences, 485849910 total letters). Two types of queries were used: a short query consisting of 20 sequences of ~400 amino acids, and a long query consisting of 20 sequences of ~900 amino acids.

Figure 1 and Figure 2 show the performance comparison of blastp on both AIX and Linux. Blastp has linear scaling up to four CPU on both operating systems. AIX showed slightly better performance than Linux. The advantage was up to 9%.

The shorter query demonstrated slightly better scalability on Linux (Figure 1), and the long query demonstrated similar scalability on both Linux and AIX (Figure 2).

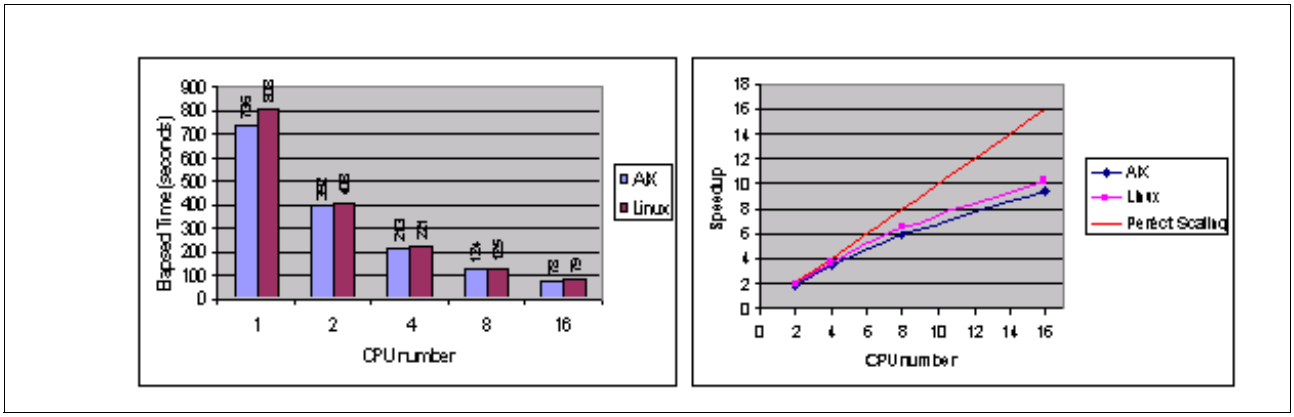

Figure 1 Performance and scalability comparison of blastp benchmark on AIX and Linux - short query

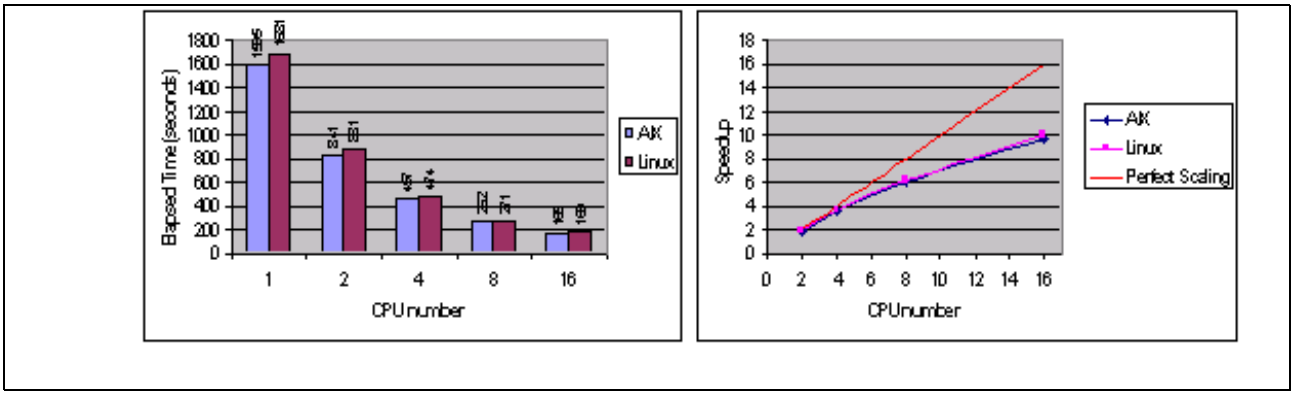

Figure 2 Performance and scalability comparison of blastp benchmark on AIX and Linux - long query

Blastn

In this study, two types of searches were used. In the short search, a query of four mouse cDNA of ~350 nucleotides was used to search against a mouse EST database (3819720 sequences, 1715033129 total letters). In the long search, a query of two *Saccharomyces cerevisiae* genomic DNA of ~1700 nucleotides was used to search against an nt database (1979572 sequences, 9422064200 total letters).

Figure 3 and Figure 4 on page 4 show the performance and scalability comparison of blastn on both AIX and on Linux. AIX showed very similar performance to Linux for both queries.

The shorter query demonstrated slightly better scalability on Linux (Figure 3), and the long query demonstrated similar scalability on both Linux and AIX (Figure 4).

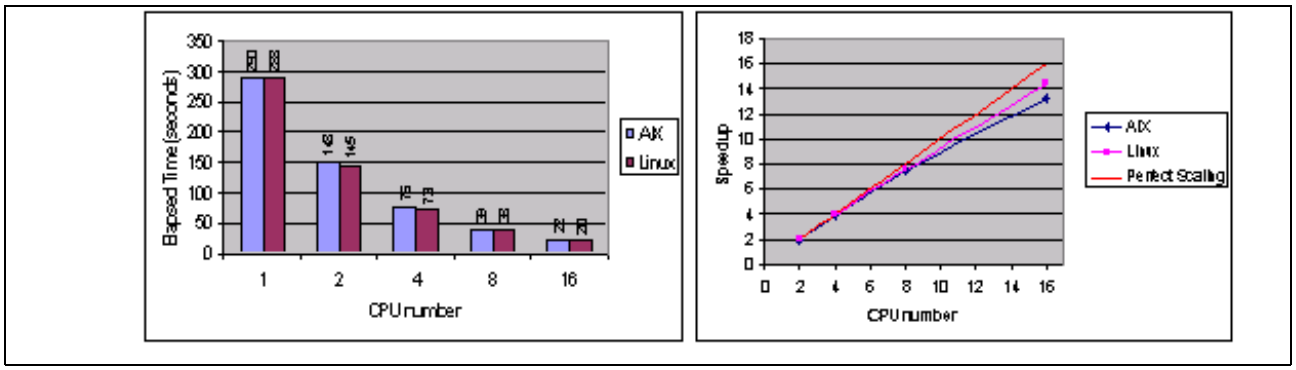

Figure 3 Performance and scalability comparison of blastn benchmark on AIX and Linux - short query

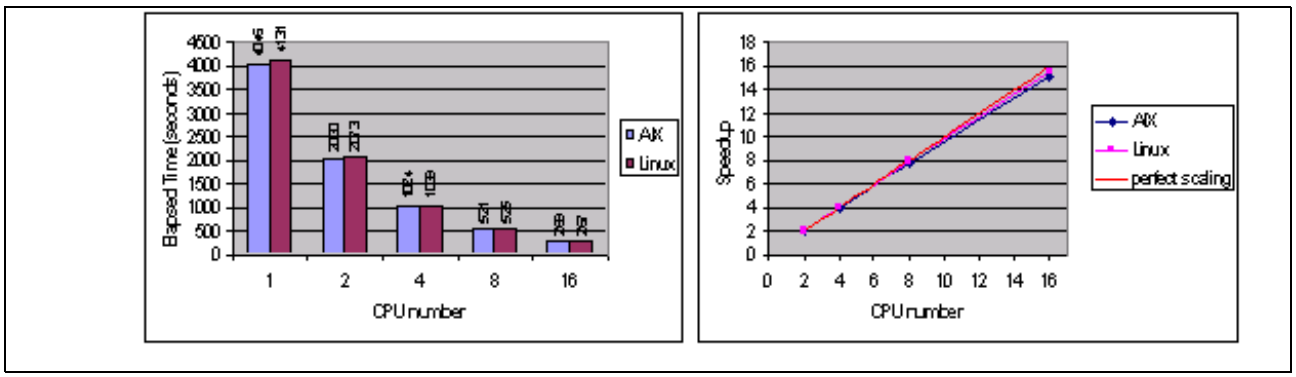

Figure 4 Performance and scalability comparison of blastn benchmark on AIX and Linux - long query

Blastx

In this study, *Drosophila* nucleotide sequences were used to search against an nr database (1508490 sequences, 485849910 total letters). Two types of queries were used: the short query was a sequence of ~ 6000 nucleotides, and the long query was a sequence of 25000 nucleotides.

Figure 5 and Figure 6 on page 5 show the performance and scalability comparison of blastx on both AIX and on Linux. AIX showed better performance than Linux, especially on the longer search. For the long query, the advantage was up to 20%.

AIX also demonstrated slightly better scalability than Linux for both test cases. AIX demonstrated up to 13% better scalability in the short query case.

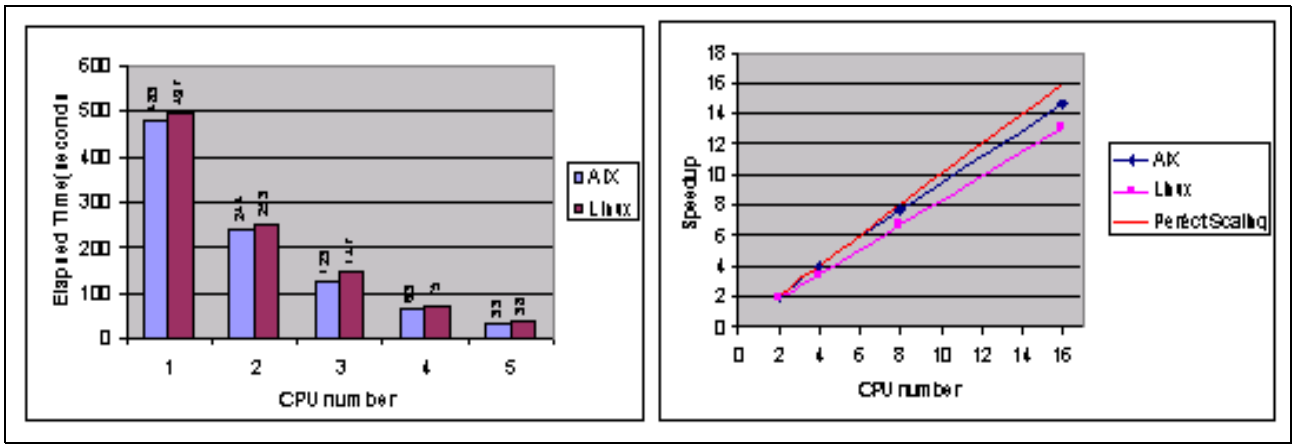

Figure 5 Performance and scalability comparison of blastx benchmark on AIX and Linux - short query

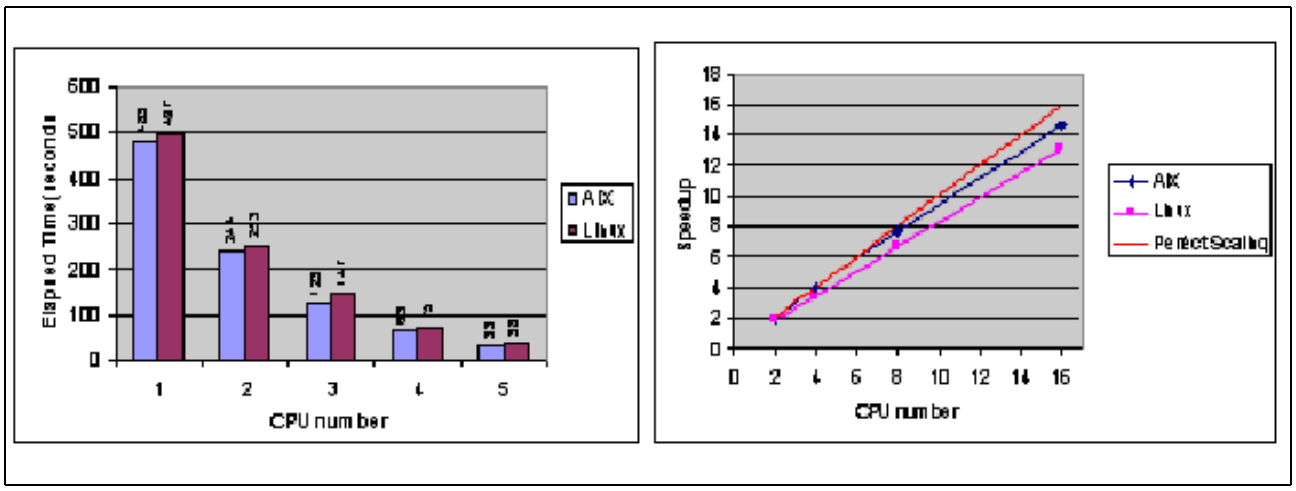

Figure 6 Performance and scalability comparison of blastx benchmark on AIX and Linux - long query

FASTA

FASTA is a widely used package for database similarity search with a high degree of sensitivity (3, 4). It is based on an alignment algorithm applicable to both nucleotide and protein sequence search for local alignment using a substitution matrix. Among several programs in the FASTA package, fasta34_t from fasta34t22b3 was used for this scalability study.

A mouse DNA sequence of 1660 nucleotides was used as a query to search against a mouse EST database (3819720 sequences, 1715033129 total letters). The benchmark results (shown in Figure 7 on page 6) indicated that Linux outperformed AIX. The advantage was up to 9%. A similar observation had been previously reported (5). The scalability on both operating systems is very similar and is almost perfectly linear.

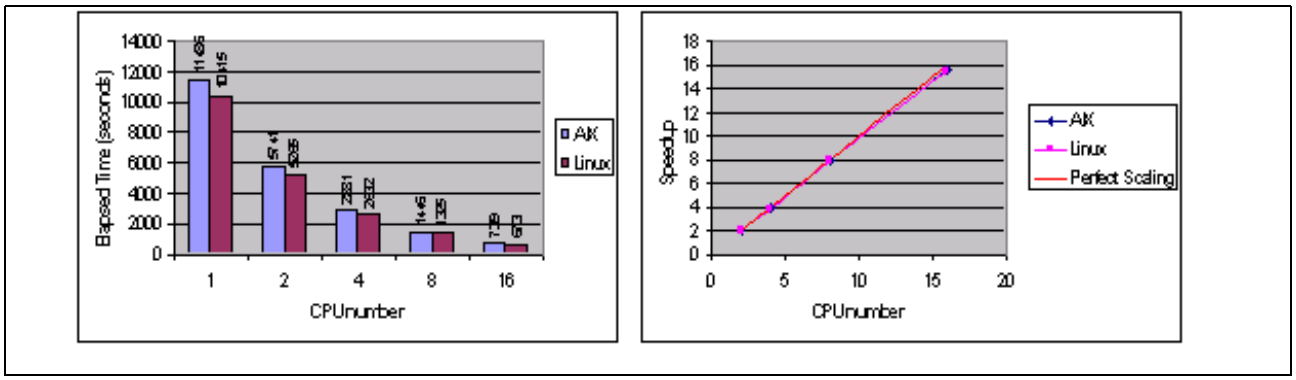

Figure 7 Performance and scalability comparison of fasta34_t on AIX and on Linux

Smith-Waterman

The Smith-Waterman program is a very sensitive database search algorithm developed by Smith and Waterman (5). In contrast to BLAST and FASTA, it looks for the optimal alignment of a query with all possible lengths, and maximizes the similarity measure. Distantly related sequences with extensive substitutions and gaps can be identified.

For this scalability study, ssearch34_t from fasta34t22b3 was used; it can use either a nucleotide as a query to search against a nucleotide database, or use a protein sequence as a query to search against a protein database.

Two test cases were used. A mouse DNA of 80 nucleotides was used to search against a mouse EST (3819720 sequences, 1715033129 total letters), and a mouse protein of 413 amino acids was used to search against the Swiss-Prot database (131418 sequences, 48074139 total letters).

The results (see Figure 8 and Figure 9 on page 7) show that AIX had slightly better performance than Linux for the DNA search. The advantage was only up to 2%. Both test cases illustrated very similar scalability on both AIX and Linux.

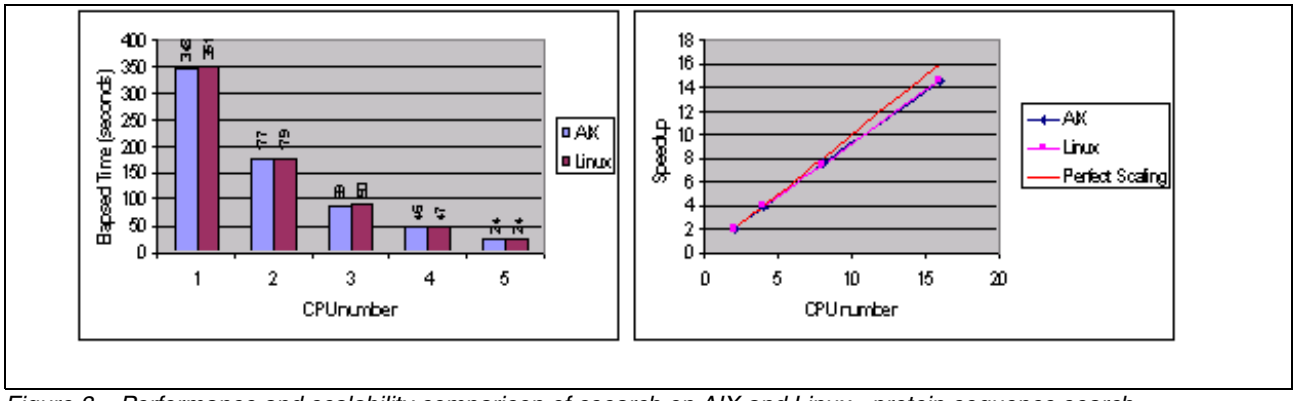

Figure 8 Performance and scalability comparison of ssearch on AIX and Linux - protein sequence search

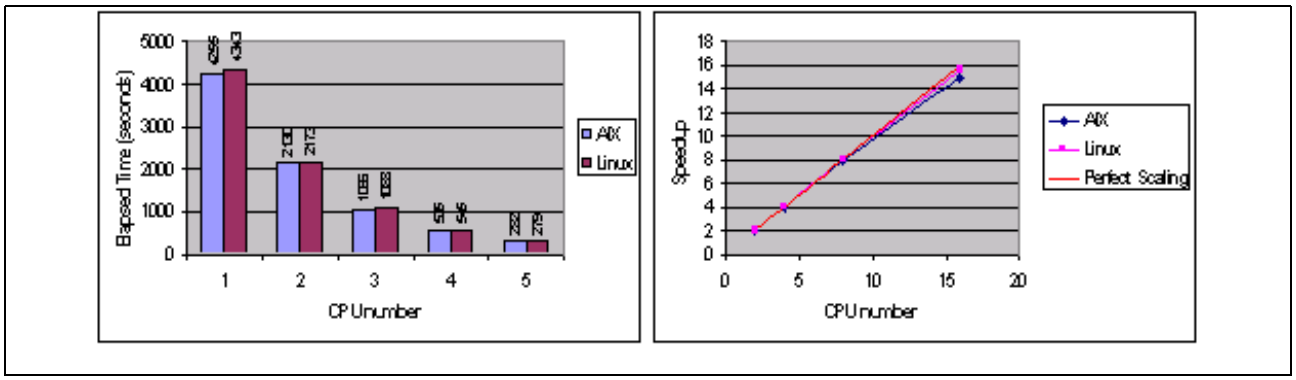

Figure 9 Performance and scalability comparison of ssearch on AIX and Linux - DNA sequence search

HMMER

HMMER is an implementation of profile hidden Markov models (profile HMMs) (7, 8). It compiles the results of a multiple alignment and generates a statistical description of a sequence family's consensus. Among several programs in the HMMER package, hmmsearch is the most computational-intensive and is implemented in parallel; hmmsearch used in this study is from HMMER 2.3.1. It uses a profile HMM as the query to search against a sequence database for additional homologues of a modeled family.

In this benchmark, we searched the Swiss-Prot database (131418 sequences, 48074139 total letters) using an HMM with 353 amino acids. The results (see Figure 10) show that AIX had a slight advantage over Linux in both runtime and scalability.

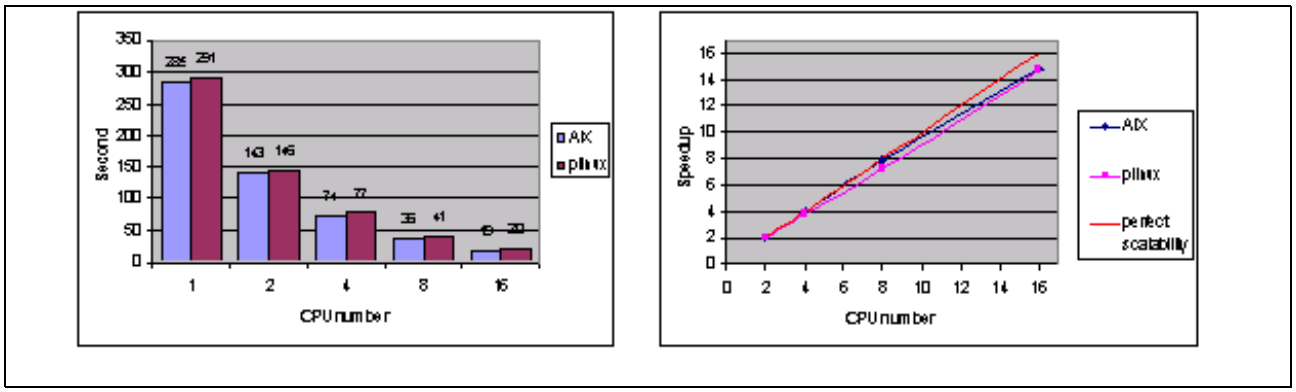

Figure 10 Performance and scalability comparison of hmmsearch on AIX and Linux

ICED

ICED (Independently Consistent Expression Discriminator) (9) is a sample classification application for high-throughput gene expression data. This application is used to build binary predictors and identify disease-relevant genes.

ICED has three functionalities: Training, Testing and Cross-validation.

- In Training, a predicting system is built based on a labeled dataset.
- In Testing, unknown samples are classified by a pre-trained system.
- Cross-validation tests the accuracy of the system.

The parameter *p* is the portion of the total number of genes that the program considers in the dynamic discriminator number selection.

Three test cases were used for benchmarks. One of the test cases was a training test of a St. Jude dataset (10), with *p* set to 0.8. Figure 11shows the performance and scalability comparison on AIX and Linux.

The other two test cases were cross-validation tests of the St. Jude dataset. Figure 12 and 13 show the benchmark results with *p* set to 0.05 and 0.1, respectively. The scalability of ICED on both operating systems was very similar. It is closer to perfect for larger jobs. It also shows that AIX achieved slightly better performance than Linux.

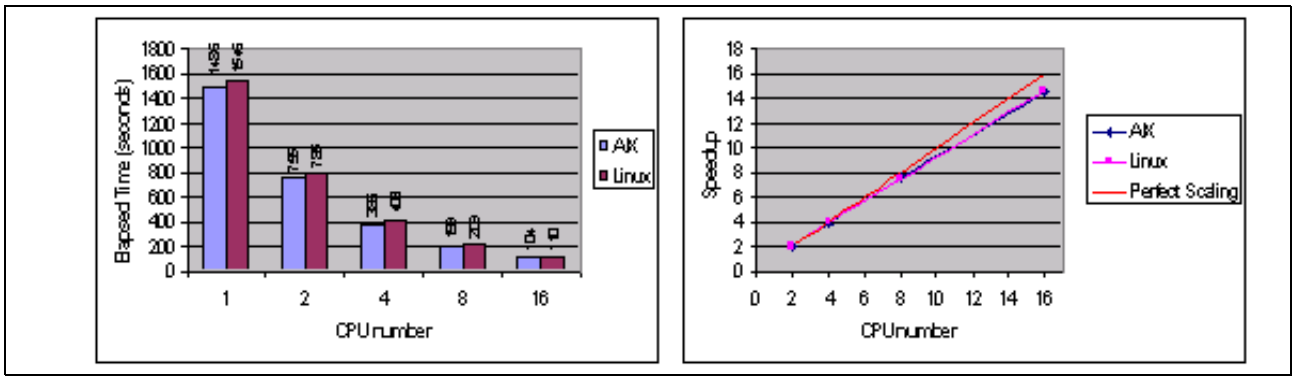

Figure 11 Performance and scalability comparison of ICED on AIX and Linux - training test of St. Jude data set, p=0.8

The other two test cases were cross-validation tests of the St. Jude dataset. Figure 12 and Figure 13 on page 9 show the benchmark results with *p* set to 0.05 and 0.1, respectively.

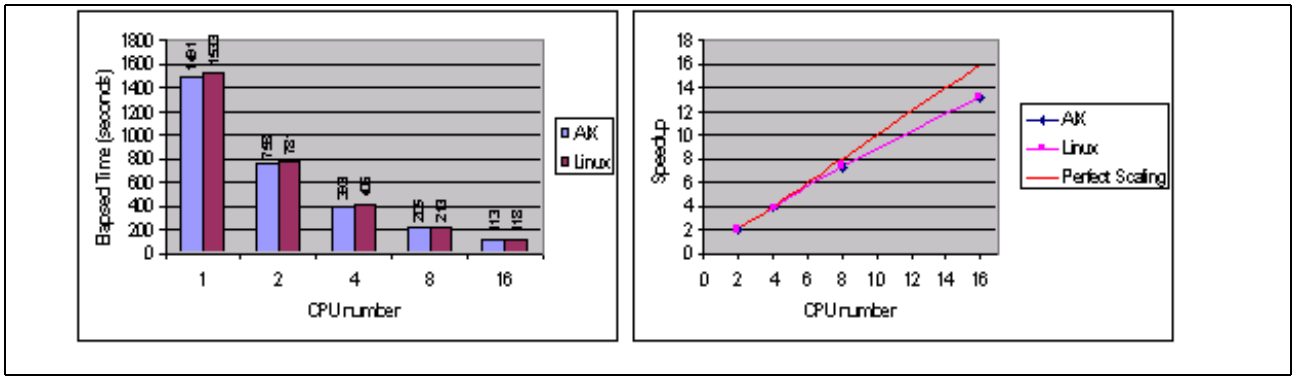

Figure 12 Performance comparison of ICED on AIX and Linux -- cross-validation test of St. Jude data set, p = 0.05

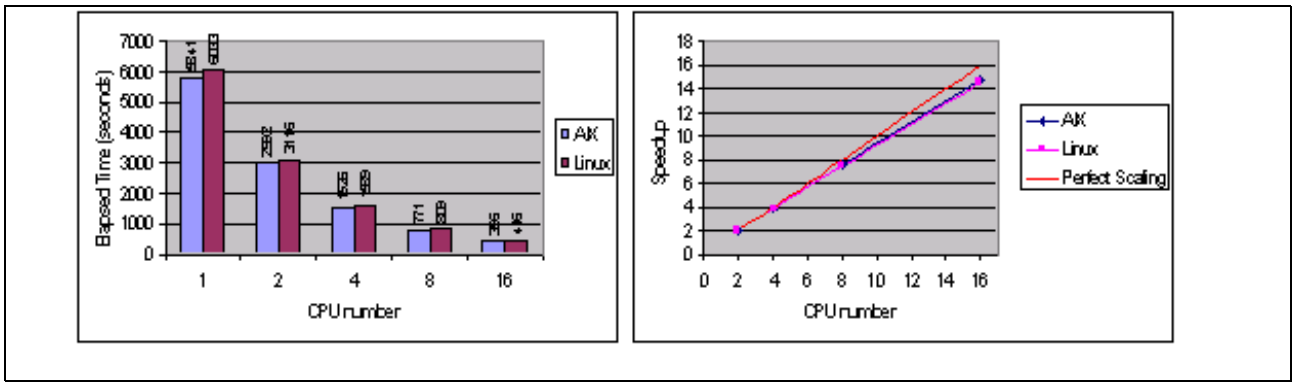

Figure 13 Scalability comparison of ICED on AIX and pLinux - cross-validation test of St. Jude data set, p=0.1

Mascot

Mascot[1] is a leading proteomics application which analyzes mass spectrometry data and identifies proteins by searching against either nucleotide or protein databases (11). Three different test cases were used for evaluation:

- Test case 1 was a Tryptic digest search of 4 ms-ms data sets.
- Test case 2 was a no enzyme search of 4 ms-ms data sets.
- Test case 3 was a Tryptic digest search of 169 ms-ms data sets with 8 variable modifications.

The database used was a MSDB database with 672745 sequences.

The benchmark results showed that test case 1 scaled up to four CPUs on both AIX and SUSE Linux (see Figure 14). This was expected, since the elapsed time for one CPU run is only 21 seconds for both operating systems.

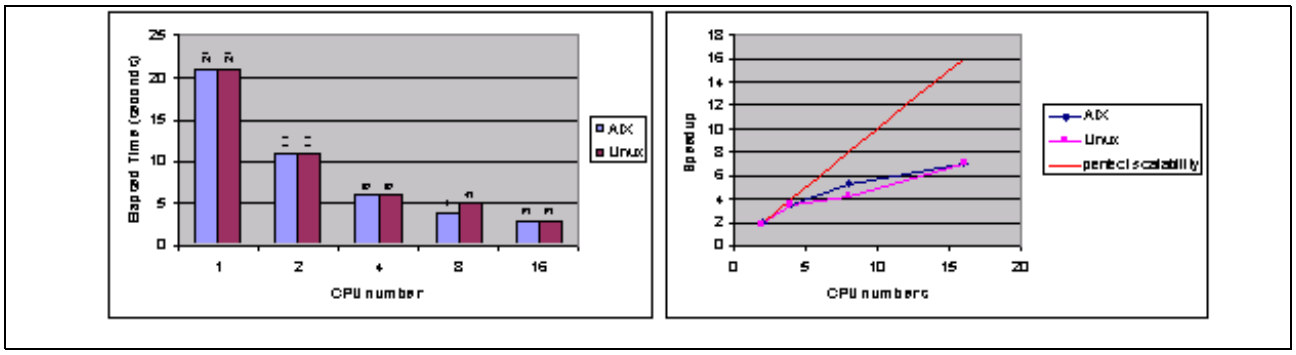

Figure 14 Performance and scalability comparison of Mascot on Linux and AIX - test case 1

Figure 15 and Figure 16 on page 10 illustrate the scaling performance for test case 2 and test case 3, respectively. Both test cases show almost linear scalability on both AIX and Linux. The performance on both systems was also almost the same.

[1] Available from MatrixScience at: http://www.matrix-science.com/.

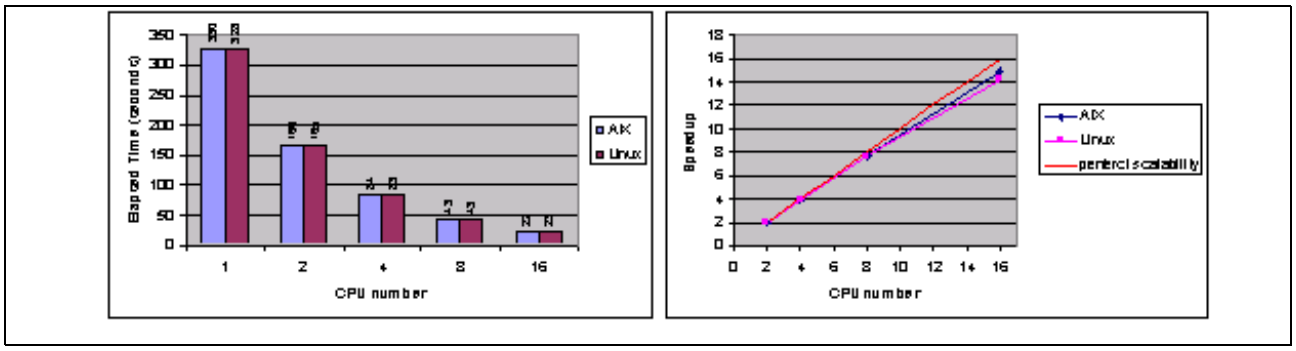

Figure 15 Performance and scalability comparison of Mascot on Linux and AIX - test case 2

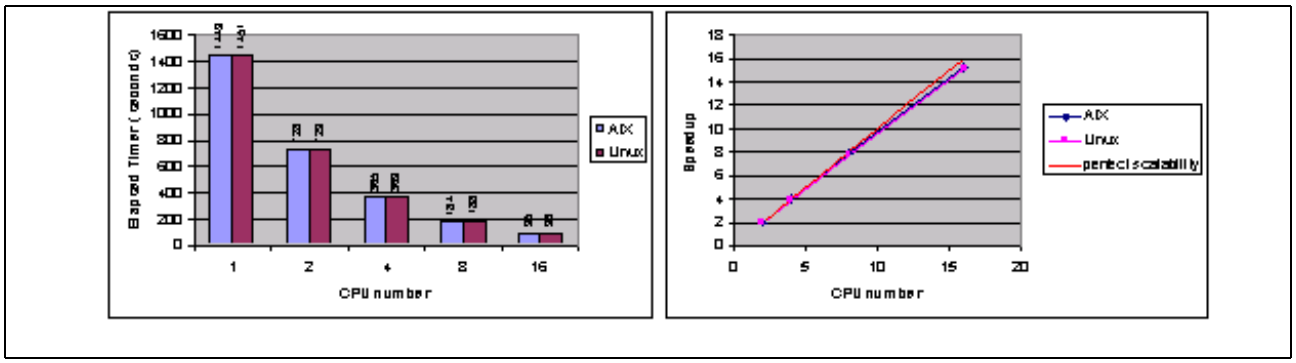

Figure 16 Performance and scalability comparison of Mascot on Linux and AIX - test case 3

Conclusion

AIX achieved better performance than Linux on most tested bioinformatics applications, except for FASTA. The performance differences were small; for most test cases, it was less than 10%. The scalability of AIX and Linux is comparable, in general. For some applications, one operating system showed slightly better scalability than the other. But neither of them consistently achieved significantly better scalability, and the difference was usually negligible.

The team that wrote this paper

Yinhe Cheng is a PhD candidate in the Department of Computer Sciences, University of Rochester, and was a Co-op with IBM during the summer of 2003. She is one of the primary designers and developers of ICED. Her research interests include bioinformatics, machine learning, and data mining.

Joyce Mak is a Software Engineer at IBM eServer pSeries Benchmark and Enablement Center. She is a certified AIX system administrator and has many years of experience managing pSeries servers for Commercial and High Performance Computing benchmarks.

Chakarat Skawratananond is a technical consultant in the IBM eServer Solutions Enablement organization, where he assists ISVs in enabling their applications for AIX and Linux on the IBM pSeries platform.

Tzy-Hwa Kathy Tzeng is a Life Sciences Software Engineering Consultant in IBM High Performance Computing Solutions Development. She received her PhD from Iowa State University and has several years of experience in the area of High Performance Computing.

Acknowledgement

We would like to thank Bob Arenburg and Matthew Davis for providing a Linux port for the tested applications.

References

1. IBM eServer Linux for pSeries: An Overview for Customers, IBM White Paper, February 2003, available at:

   http://www-1.ibm.com/linux/files/linux_pseries.pdf

2. Altschul, S. F., Fish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. J. Mol. Bio. 215: 403-410.
3. Lipman, D. J., and Pearson, W. R. 1985. Rapid and sensitive protein similarity searches. Science 227:1435-1441.
4. Pearson, W. R., and Lipman, D. J. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85:2444-2448.
5. Smith T.F. and Waterman M.S. (1981) Identification of common molecular sub-sequences. Journal of Molecular Biology, 147:195-197.
6. Performance Comparison of Scientific Applications on AIX and Linux for PowerPC®, IBM Redpaper, REDP-3692-00, March, 2003.
7. Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. 1994. J. Mol. Biol., 235:1501-1531.
8. Eddy, S. R. 1996. Curr. Opin. Struct. Biol., 6:361-365.
9. Bijlani, R., Cheng, Y., Pearce, D. A., Brooks, A. I. and Ogihara, M. 2002. Prediction of biologically significant components from microarray data: Independently Consistent Expression Discriminator (ICED). Bioinformatics 18:1-9.
10. Yeoh E J, Ross ME, Shurtleff SA, Williams Wk, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling Mv, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui Ch, Evans WE, Naeve C, Wong L, and Downing JR. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell, 1(2):109-10.
11. Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. 1999. Probability-based protein identification by searching sequence database using mass spectrometry data. Electrophoresis 20:3551-3567.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:
This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Send us your comments in one of the following ways:

- Use the online **Contact us** review redbook form found at:
  ibm.com/redbooks
- Send your comments in an Internet note to:
  redbook@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYJ Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX®	IBM®	Redbooks (logo)™
AIX 5L™	PowerPC®	VisualAge®
@server®	POWER™	@server®
@server®	POWER4™	
pSeries®	Redbooks™	

The following terms are trademarks of other companies:

Intel and Intel Inside (logos) Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.