Parallel taxonomic classification algorithm for metagenomic sequences

Le Van Vinh, Tran Van Hoai, Duong Ngoc Hieu, Bui Xuan Giang, Tran Van Lang, Thoai Nam

Bach Khoa University

March 1, 2017

Overview

1. Introduction
2. Related works
3. Proposed method
4. Results
5. Conclusion

Metagenomics

- Known as *environmental genomics* or *community genomics*
- Study of uncultured microorganisms
- Contribute to advances in many fields, e.g., Earth sciences, life sciences, bioenergy, biotechnology, agriculture.

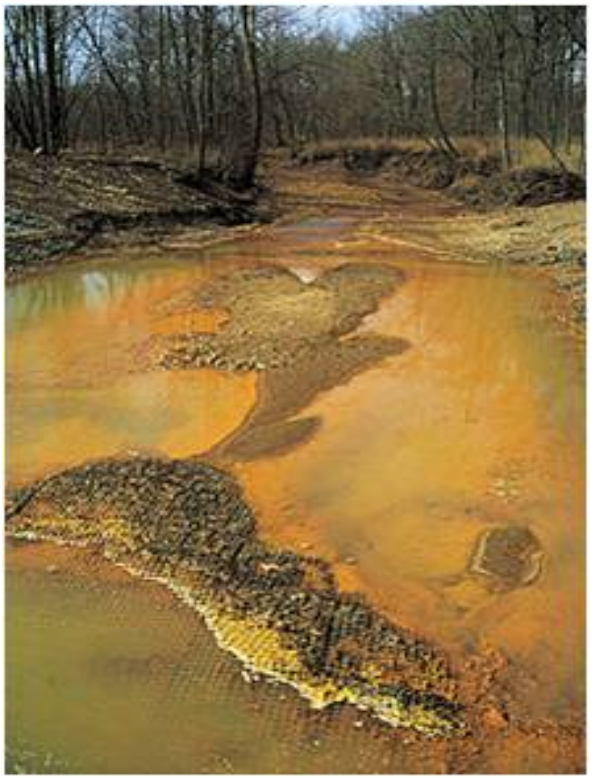
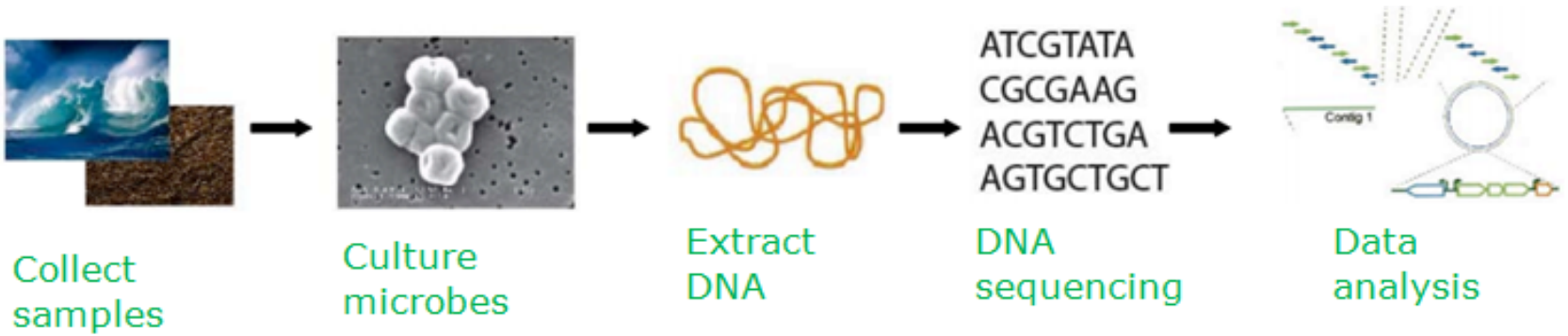


Figure: Acid drainage(wikipedia.org)

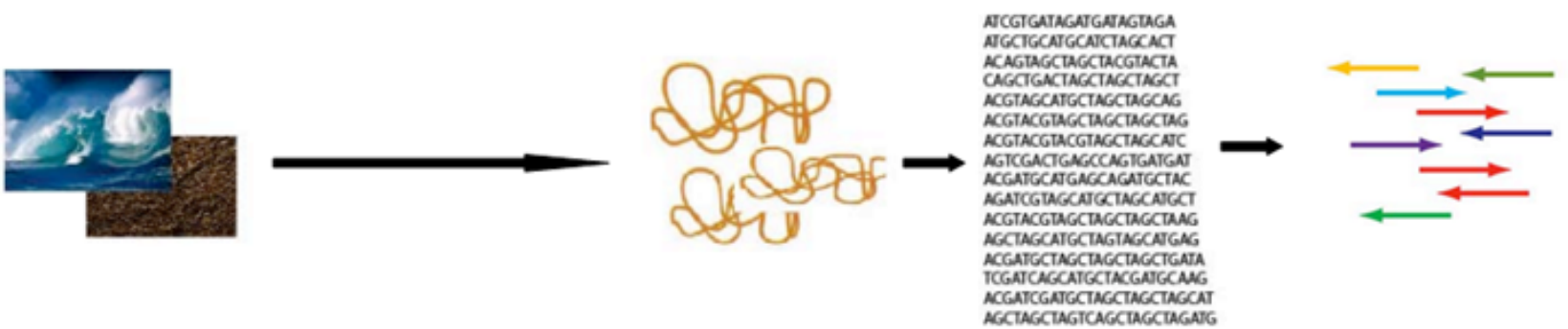
Genomic vs. Metagenomics

Problem: more than 99% microbes cannot be cultured in the laboratory

Genomics vs. Metagenomics



Problem: more than 99% microbes cannot be cultured in the laboratory



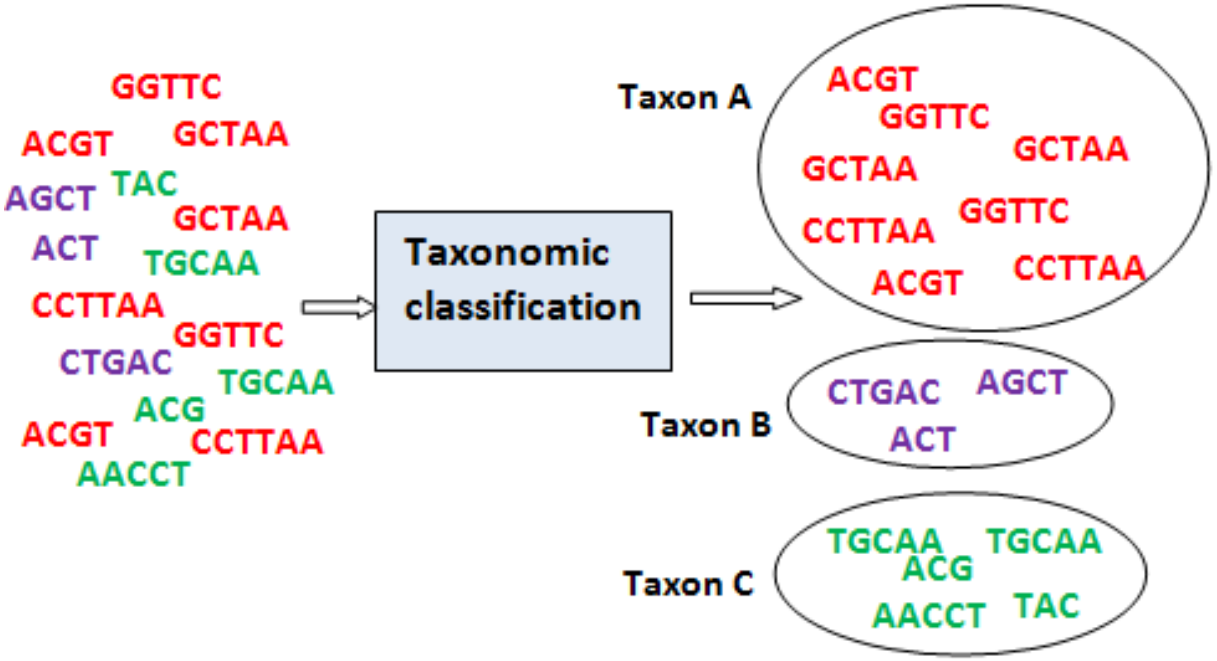
Problem: lose information about which genes/sequences belong to which genomes/microbes

Source: Gianoulis, Harvard university, US

Taxonomic classification of metagenomic reads

Problem's definition

The taxonomic classification aims to group reads into bins and determines phylogenetic relationships between them and known taxa.

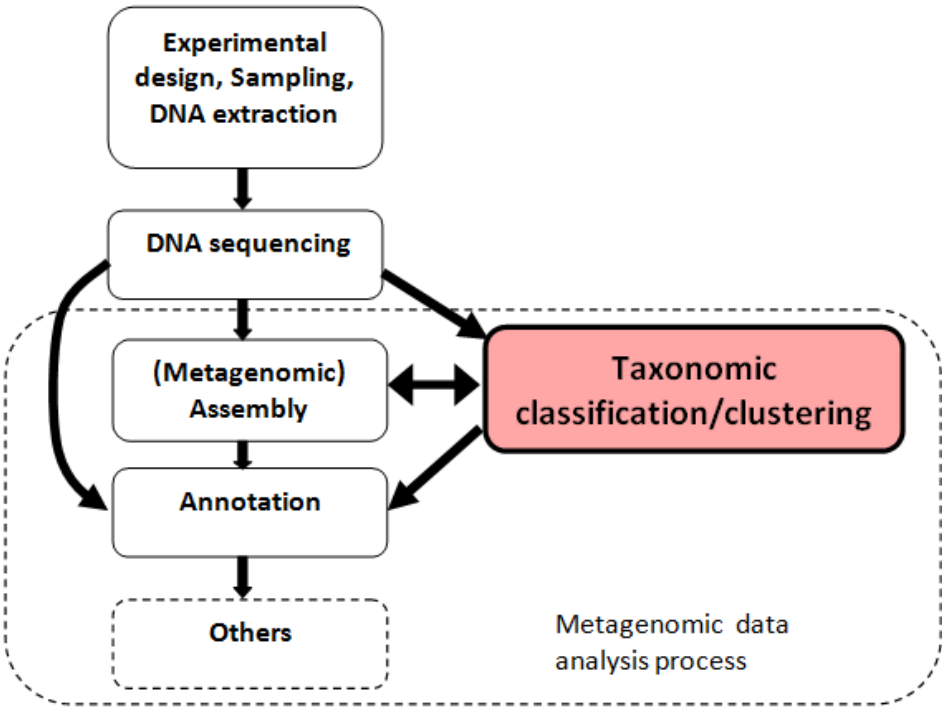


Why classification problem?

(1) The primary goals of metagenomic studies [2]

- Who is out there? (taxonomic content, abundance level?)
- What are they doing?
- How do they compare?

(2) An important step in a metagenomic project [1]



Research challenges

Short read length

- Genomic signatures are less preserved in DNA read with length $< 1000bp$ [3]
- RAIphy (2011)[4]: 32% - 36% (Accuracy) for reads with the length of 100bp

Research challenges

Short read length

- Genomic signatures are less preserved in DNA read with length $<$ 1000bp [3]
- RAIphy (2011)[4]: 32% - 36% (Accuracy) for reads with the length of 100bp

Large amount of data

- Require large computational costs
- Work with a huge amount of reference database (GenBank (12/2016): $\approx$ 200.000.000 sequences with $\approx 2 \times 10^{13}$ bases)

Related works

- **Composition-based methods**
  - Using genomic signatures (e.g., oligonucleotide frequencies, GC-content)
  - TACOA [6], AKE [7])
  - are fast, but difficult to analyze short reads
- **Homology-based methods**
  - Basing on the similarity between sequences
  - MEGAN [2], CARMA3 [9], MetaCluster-TA [10]
  - work well with both short and long reads, but much computational expense

Related works

Metagenomic applications are based on high-performance computing techniques

- **Map-reduce framework**: MrMC-MinH [11],
- **GPU, multi-core-CPU**: Parallel-META [12]
- **MPI**: mpiBlast [13]

Related works

Our previous works

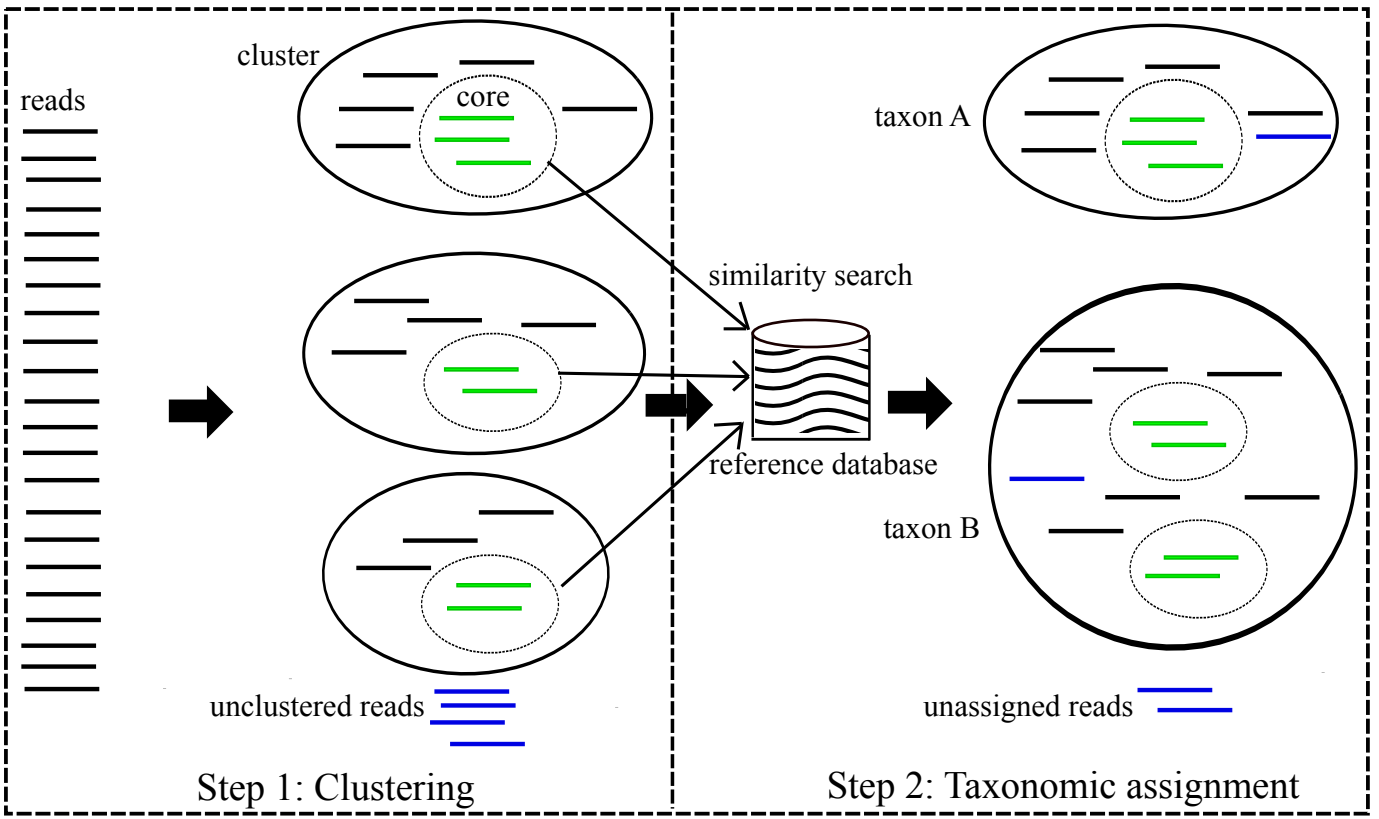


Figure: Classification process of SeMeta (Vinh *et al* (2016) [8])

ParSeMeta

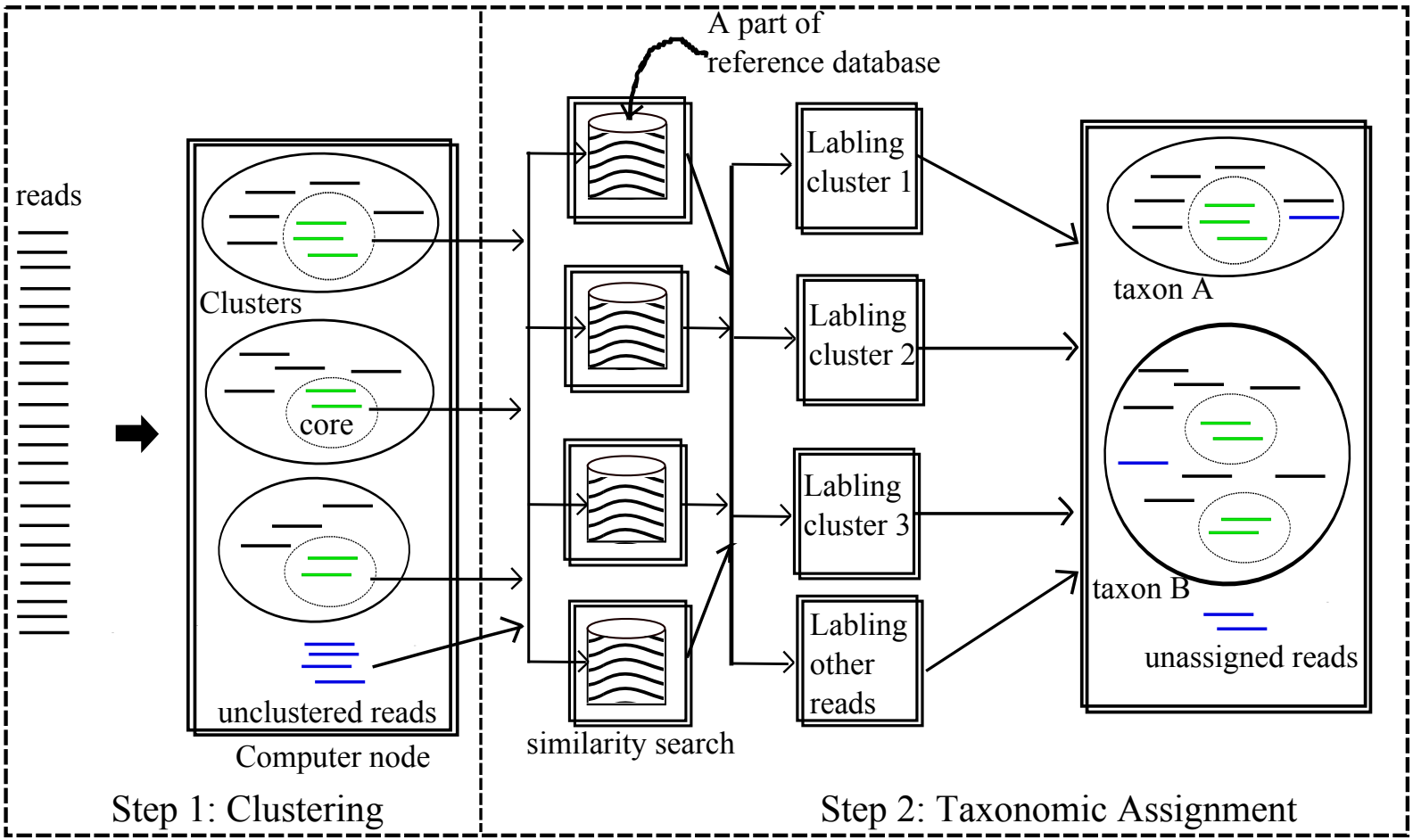


Figure: Classification process of ParSeMeta

Experiment setup

- Test on two physical machines (12 CPUs, 120G RAM, and 100GB disk storage)
- Three cases (number of cores, number of virtual machines, memory sizes)

ParSeMeta

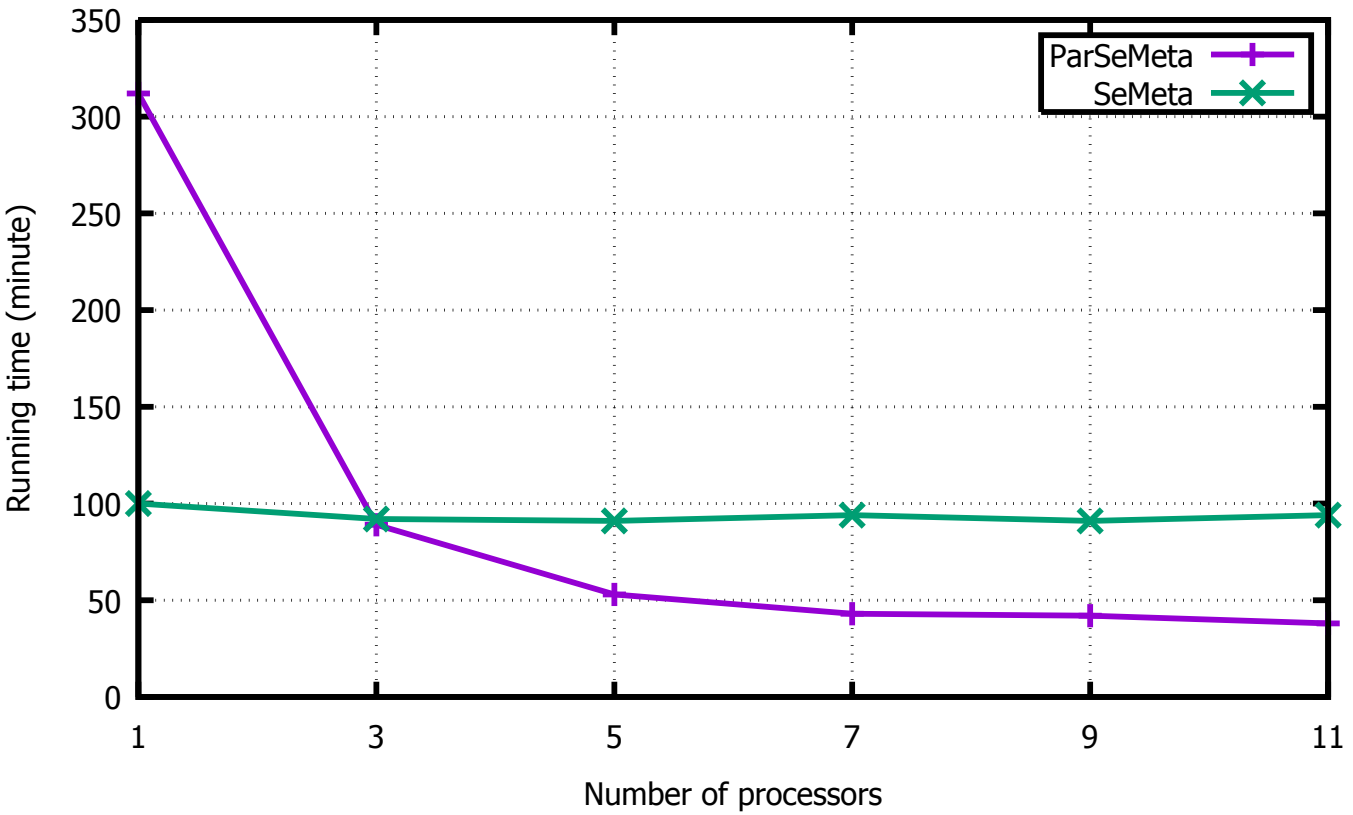


Figure: The performance of ParSeMeta and SeMeta with different numbers of core

ParSeMeta

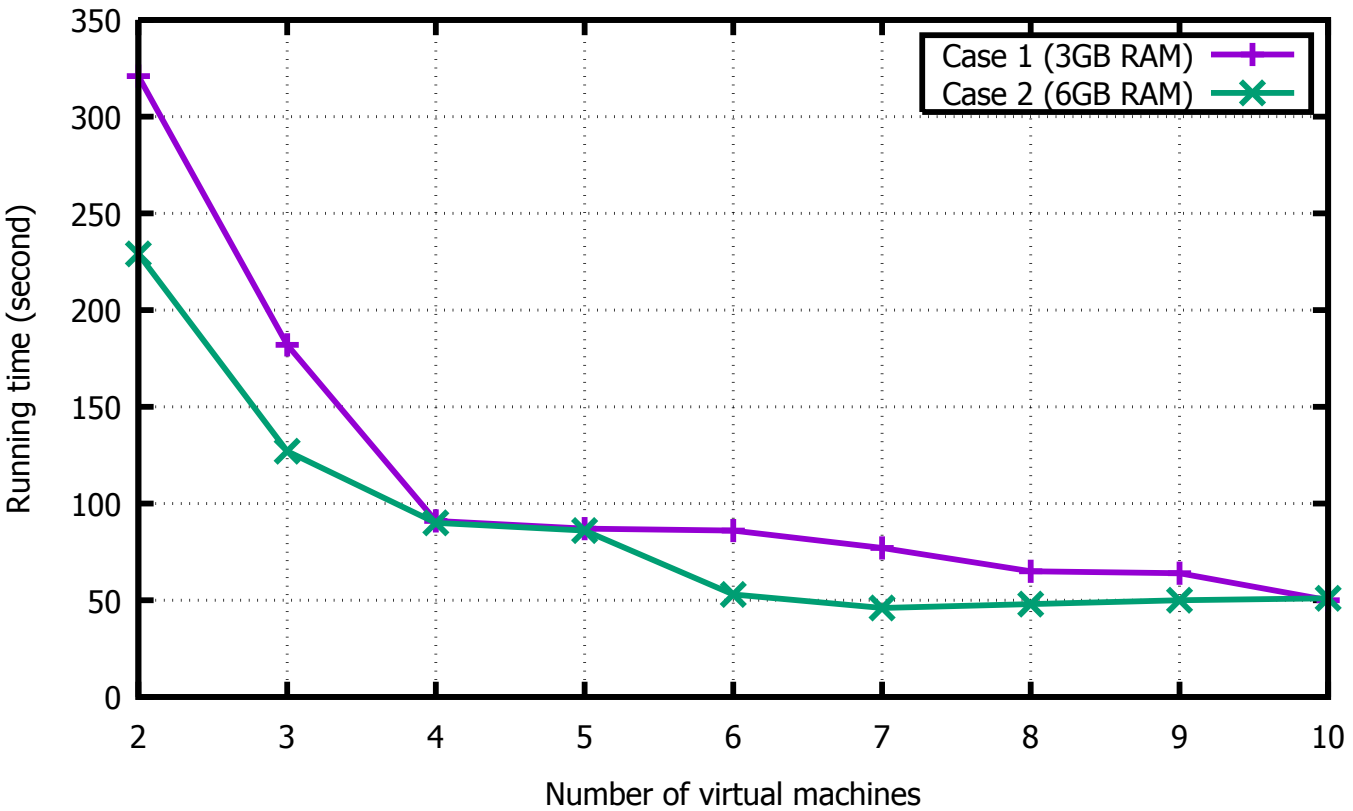


Figure: The performance of ParSeMeta with different numbers of virtual machines, with cases of using 3GB RAM and 6GB RAM

The classification quality

Dataset ds1		Species level	Genus level	Family level	Order level
SeMeta	*Sen.*	**42.76%**	42.76%	99.71%	99.71%
	Pre.	**42.88%**	42.88%	100%	100%
ParSeMeta	*Sen.*	N/A	**99.71%**	99.71%	99.71%
	Pre.	N/A	**100%**	100%	100%
Dataset ds2					
SeMeta	*Sen.*	24.72%	30.24%	61.94%	61.94%
	Pre.	39.91%	30.34%	100%	100%
ParSeMeta	*Sen.*	24.72%	30.24%	61.94%	61.94%
	Pre.	39.91%	30.34%	100%	100%
Dataset ds3					
SeMeta	*Sen.*	**46.69%**	64.84%	64.84%	64.84%
	Pre.	**67.09%**	93.16%	93.16%	93.16%
ParSeMeta	*Sen.*	23.64%	64.84%	64.84%	64.84%
	Pre.	16.45%	93.16%	93.16%	93.16%

- Propose a parallel algorithm utilizing the advantages of high performance computing system
- ParSeMeta is able to reduce much computational time, still keeps classification quality
- Future works: apply on large-scale metagenomic datasets, predicting execution time of the algorithm with different settings (parameters of algorithm, allocated resources)

References

[1]. Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics-a guide from sampling to data analysis. Microb Inform Exp, 2(3), 1-12.
[2]. Huson D. H. *et al* (2011). Integrative analysis of environmental sequences using MEGAN4. Genome Res. 21 (9).
[3]. Deschavanne, P. *et al* (2000). Genomic signature is preserved in short DNA fragments. In Bio-Informatics and Biomedical Engineering, 2000. Proceedings. IEEE International Symposium on (pp. 161-167). IEEE.
[4]. Nalbantoglu, O. U. *et al* (2011). RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC bioinformatics, 12(1), 41.
[5]. Eisen J. A.. Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes. PLoS Biol., vol. 5, no. 3, 2007.
[6]. Diaz, N. N.et al (2009). TACOATaxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC bioinformatics, 10(1), 56.
[7]. Langenkmper, D., Goesmann, A., and Nattkemper, T. W. (2014). AKE-the Accelerated k-mer Exploration web-tool for rapid taxonomic classification and visualization. BMC bioinformatics, 15(1), 384.

References

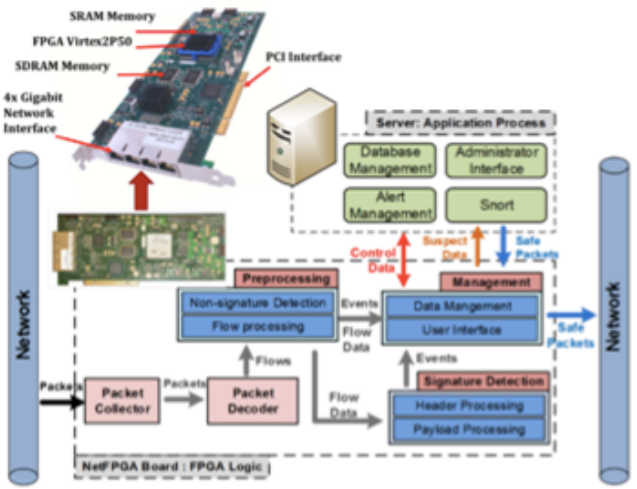
[8]. Vinh V. L., Lang V. T., and Hoai V. T. (2016). A novel semi-supervised algorithm for the taxonomic assignment of metagenomic reads. *BMC Bioinformatics*, 17 (22).
[9]. Gerlach *et al* (2011). Taxonomic classification of metagenomic shotgun sequences with carma3. *Nucleic acids research*, 39 (14).
[10]. Yi *et al* (2014). Metacluster-ta: taxonomic annotation for metagenomic databased on assembly-assisted binning. *BMC Genomics*, 15.
[11]. Rasheed *et al* (2013). A map-reduce framework for clustering metagenomes. In: *Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW)*.
[12]. Su *et al* (2012). Parallel-meta: efficient metagenomic data analysis based on high-performance computation. *BMC Systems Biology*, 6 (1).

[13]. Darling *et al* (2003). The design, implementation, and evaluation of mpiblast.

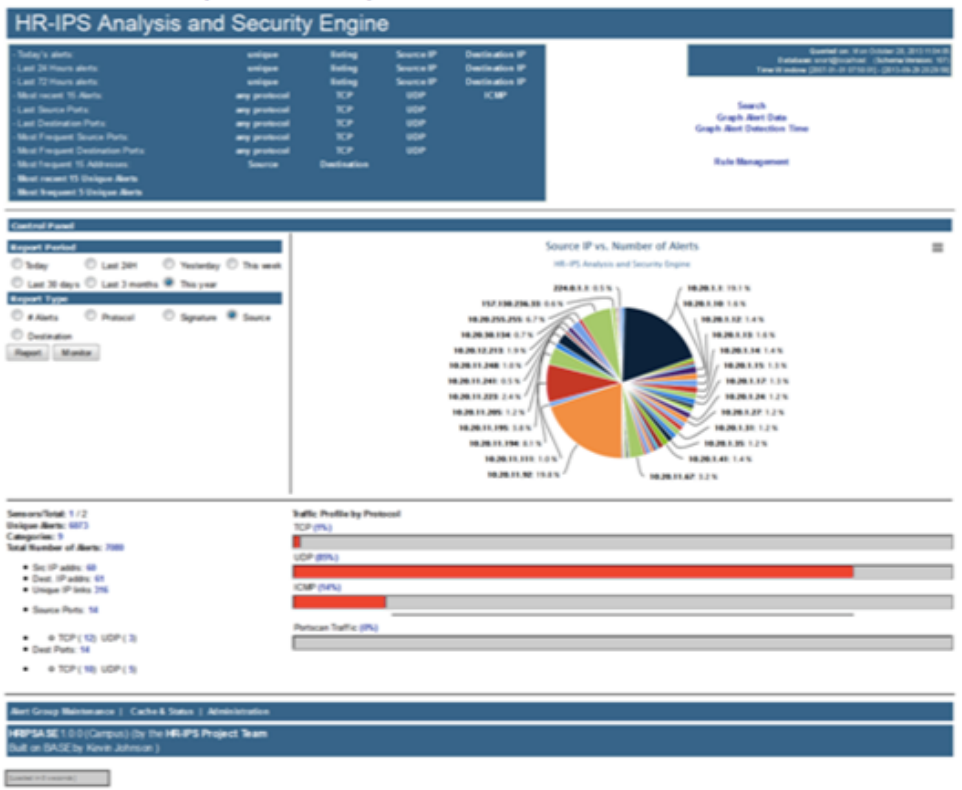
Proceedings of ClusterWorld.

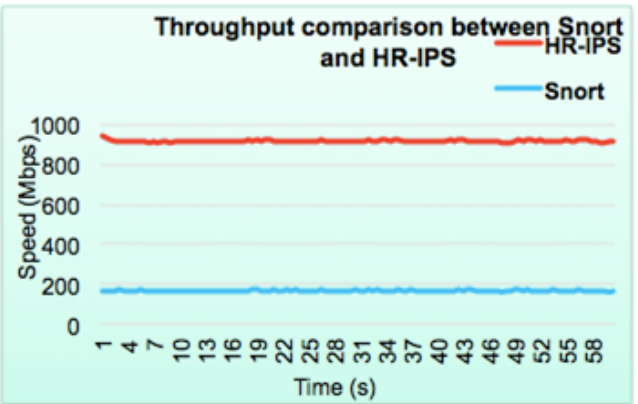
Applying Reconfigurable Computing for Internet Security

– HR-IPS: HCMUT Reconfigurable Intrusion Prevention System



- HR-IPS meets line rate full duplex **Gbps network** link without any drops.
- **5x better** throughput compare to bare software solution (Snort IPS).





Low-power wireless water quality monitoring system

- Water quality monitoring system for shrimp farming
- Energy harvesting
- Low-power network protocol
- Data management