Report Summary: A Survey of High Performance Computing (HPC) Infrastructure in Thailand

นางสาว กัญญ์ธรรม ถังทอง

รหัสนักศึกษา 63010054

17 กันยายน พ.ศ. 2567

บทคัดย่อ

เนื้อหาในรายงานฉบับนี้ได้สรุปความในเอกสารทางวิชาการหัวข้อ A Survey of High-Performance Computing (HPC) Infrastructure in Thailand เขียนโดย รองศาสตราจารย์ ดร.วรา วราวิทย์ และ ผศ.ดร.ศุภกิจ พฤกษอรุณ โดยผู้เขียนทั้งสองได้เน้นการนำเสนอรายละเอียดความเป็นมาของโครงสร้างพื้นฐานของเทคโนโลยีการประมวลผลประสิทธิภาพสูง (High-Performance Computing: HPC) ทั้งในภาคเอกชนและรัฐบาล ไปจนถึงหน่วยงานและสถาบันการวิจัยระดับชาติอย่าง ศูนย์ทรัพยากรคอมพิวเตอร์เพื่อการคำนวณขั้นสูง (NSTDA Supercomputer Center: ThaiSC) เพื่อให้ผู้ที่กำลังศึกษาหรือกำลังจะศึกษาค้นคว้าวิจัยในสายงานนี้มีความเข้าใจภาพรวมของ HPC ในประเทศไทย รวมถึงผู้แต่งยังได้นำเสนอแนวทางการประยุกต์ใช้ HPC ในภาคอุตสาหกรรมตั้งแต่อดีตจนถึงปัจจุบัน โดยเฉพาะเรื่องการถ่ายโอนข้อมูลขนาดใหญ่ผ่าน Inter-University Network หรือ UniNet อันเป็นเครือข่ายเพื่อการเป็นศูนย์กลางด้านการศึกษาวิจัยระดับชาติ ซึ่งอยู่ภายใต้การกำกับดูแลโดยสำนักงานบริหารเทคโนโลยีสารสนเทศ (Office of Information Technology Administration for Educational Development)

สารบัญ

1 บทนำ (Introduction)

ภาพรวมของแบบสำรวจนี้ อธิบายถึงการประยุกต์ใช้เทคโนโลยี HPC ในประเทศไทย โดยเทคโนโลยีปัจจุบันได้ทำให้องค์กรหรือหน่วยงานต่างๆต้องหันมาพึ่งพาความสามารถในการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data) ซึ่งการดำเนินงานเหล่านี้หมายถึงต้องมีการจัดการ วัฏจักรการทำงาน (Conputing Cycle), การจัดเก็บข้อมูล (Storage Space), และแบนด์วิดท์ของเครือข่าย (Network Bandwidth) โดยความซับซ้อนและขนาดของการจัดการปัญหาเหล่านี้จะต้องพึ่งพาความสามารถและประสิทธิภาพของระบบ HPC ซึ่งถือเป็นโครงสร้างพื้นฐานทางดิจิทัล (Digital Infrastructure) ในการวิจัย

ระบบการประมวลผลขนาดใหญ่ที่ตั้งอยู่ในศูนย์กลางฐานข้อมูล (Data Center) จะมีความสามารถสองด้านในการจุข้อมูลและสมรรถนะในการคำนวณและประมวลผล กว่า 5 ทศวรรษที่มีการพัฒนาระบบ HPC ได้มีวิวัฒนาการของระบบวิทยาการคำนวณอย่างคอมพิวเตอร์เพื่อการประมวลผล ออกมาเป็นประเภทดังต่อไปนี้

- เมนเฟรมคอมพิวเตอร์ (Mainframes)

  เมนเฟรมคอมพิวเตอร์ คือคอมพิวเตอร์ขนาดใหญ่ที่มีพลังการประมวลผลสูง ซึ่งออกแบบมาเพื่อรองรับงานคำนวณที่ซับซ้อนและสามารถทำงานได้อย่างมีประสิทธิภาพสูงแม้จะมีการใช้งานจากผู้ใช้หลายคนพร้อมกัน เมนเฟรมมักถูกใช้ในองค์กรขนาดใหญ่ เช่น ธนาคาร หน่วยงานรัฐบาล หรือบริษัทที่มีข้อมูลและงานที่ต้องจัดการในปริมาณมาก ถึงแม้ว่าปัจจุบันจะมีคอมพิวเตอร์ประเภทอื่นๆ ที่ได้รับความนิยมมากขึ้น แต่เมนเฟรมยังคงมีบทบาทสำคัญในองค์กรขนาดใหญ่ที่ต้องการระบบที่เสถียรและมีความปลอดภัยสูงในการจัดการข้อมูลและงานที่สำคัญ

- เมนเฟรมคอมพิวเตอร์ขนาดเล็ก (Minicomputers)

  เมนเฟรมคอมพิวเตอร์ขนาดเล็ก เป็นคอมพิวเตอร์ที่มีขนาดเล็กกว่าเมนเฟรม แต่ใหญ่กว่าไมโครคอมพิวเตอร์ (เช่น คอมพิวเตอร์ตั้งโต๊ะหรือแล็ปท็อป) ถูกพัฒนาขึ้นมาในช่วงทศวรรษ 1960 เพื่อเป็นเครื่องคอมพิวเตอร์สำหรับองค์กรขนาดกลางที่ต้องการคอมพิวเตอร์สำหรับงานประมวลผลที่มีขนาดไม่ใหญ่มากนัก แต่ยังต้องการประสิทธิภาพและความสามารถในการจัดการข้อมูลและงานที่ซับซ้อนได้ แม้ว่ามินิคอมพิวเตอร์จะเป็นที่นิยมในอดีต โดยในปัจจุบันการพัฒนาของเทคโนโลยี ทำให้คอมพิวเตอร์ส่วนบุคคล (PC) และเซิร์ฟเวอร์ (servers) มีราคาถูกลงและมีประสิทธิภาพสูงขึ้นมาก จึงทำให้การใช้งานมินิคอมพิวเตอร์ลดลง

- คอมพิวเตอร์แบบเวกเตอร์ (Vector computers)

  คอมพิวเตอร์แบบเวกเตอร์ เป็นคอมพิวเตอร์ประเภทหนึ่งที่ออกแบบมาเพื่อประมวลผลข้อมูลแบบขนานในลักษณะที่ทำให้สามารถทำงานกับชุดข้อมูลขนาดใหญ่ (หรือเวกเตอร์) ได้พร้อมกันในเวลาเดียวกัน ซึ่งมีประโยชน์อย่างยิ่งในงานที่ต้องใช้การคำนวณทางคณิตศาสตร์ที่ซับซ้อนและมีจำนวนมาก เช่น การจำลองการไหลของของไหล (fluid dynamics), การประมวลผลกราฟิก, การคำนวณโมเลกุลทางเคมี และการประมวลผลข้อมูลเชิงวิทยาศาสตร์และวิศวกรรม เป็นต้น โดยในปัจจุบัน แม้ว่าเทคโนโลยี

คอมพิวเตอร์แบบเวกเตอร์จะไม่ถูกใช้อย่างแพร่หลายในปัจจุบันเหมือนในอดีต แต่หลักการของการประมวลผลแบบเวกเตอร์ยังคงเป็นพื้นฐานของการออกแบบฮาร์ดแวร์สมัยใหม่ โดยเฉพาะใน GPU (Graphics Processing Unit) ที่ใช้หลักการเดียวกันในการประมวลผลกราฟิกและงานคำนวณทางคณิตศาสตร์

- ซูเปอร์คอมพิวเตอร์ (Supercomputers)

  ซูเปอร์คอมพิวเตอร์ คือ คอมพิวเตอร์ที่มีประสิทธิภาพในการประมวลผลสูงที่สุดเมื่อเทียบกับคอมพิวเตอร์ประเภทอื่น ๆ โดยถูกออกแบบมาเพื่อทำงานประมวลผลที่ซับซ้อนและต้องการกำลังประมวลผลสูงอย่างมาก เช่น การจำลองสถานการณ์ การคำนวณเชิงวิทยาศาสตร์ และการวิเคราะห์ข้อมูลขนาดใหญ่ เป็นต้น โดยจุดเด่นของซูเปอร์คอมพิวเตอร์คือ สามารถประมวลผลงานได้ในระดับพันล้านล้านคำสั่งต่อวินาที (FLOPS: Floating Point Operations Per Second) ซึ่งทำให้สามารถจัดการกับงานที่ซับซ้อนได้รวดเร็ว โดยในปัจจุบันนั้น ซูเปอร์คอมพิวเตอร์เป็นเครื่องมือสำคัญในงานวิจัยและการพัฒนาทางวิทยาศาสตร์ เทคโนโลยี และอุตสาหกรรมทั่วโลก

- คอมพิวเตอร์แบบขนานขนาดใหญ่ (Massively Parallel Computers)

  คอมพิวเตอร์แบบขนานขนาดใหญ่ เป็นคอมพิวเตอร์ที่ถูกออกแบบมาเพื่อประมวลผลหลายงานพร้อมกันโดยใช้หน่วยประมวลผลหลายตัว (Processors) ที่ทำงานในรูปแบบการประมวลผลแบบขนาน (Parallel Processing) ซึ่งช่วยเพิ่มความสามารถในการคำนวณที่รวดเร็วและมีประสิทธิภาพสูง ตัวอย่างของคอมพิวเตอร์แบบขนานขนาดใหญ่ คือ

  - IBM Blue Gene เป็นโครงการพัฒนาซูเปอร์คอมพิวเตอร์ที่เริ่มต้นโดย IBM ซึ่งมีวัตถุประสงค์เพื่อสร้างซูเปอร์คอมพิวเตอร์ที่มีประสิทธิภาพสูง แต่มีการใช้พลังงานน้อยกว่าระบบทั่วไป โครงการนี้ประกอบด้วยหลายรุ่นที่พัฒนาออกมา ได้แก่ Blue Gene/L (ประมวลผลสูงถึง 360 TFlop/s), Blue Gene/P (ประมวลผลสูงถึง 1 PFlop/s) และ Blue Gene/Q (ประมวลผลสูงถึง 20 PFlop/s) ซึ่งเป็นรุ่นที่พัฒนาต่อมาแต่ละรุ่น เพื่อเพิ่มประสิทธิภาพของการประมวลผลแบบขนาน (Parallel Processing)
  - Tianhe-2 หรือ รู้จักกันในชื่อ ”Milky Way 2” เป็นซูเปอร์คอมพิวเตอร์ของจีน ซึ่งเคยเป็นซูเปอร์คอมพิวเตอร์ที่เร็วที่สุดในโลกหลายปีติดต่อกัน (2013-2015) โดยถูกพัฒนาโดยมหาวิทยาลัยเทคโนโลยีป้องกันประเทศแห่งชาติจีน (National University of Defense Technology: NUDT) โดย Tianhe-2 มีความสามารถในการประมวลผลสูงถึง 33.86 PFlop/s ในการทดสอบ LINPACK ซึ่งเป็นไลบรารี่สำหรับการทดสอบประสิทธิภาพของ HPC ในการแก้ปัญหา Linear Algebra พัฒนาขึ้นเมื่อปี 1970 โดย Jack Dongarra, Jim Bunch, Cleve Moler และ Gilbert Stewart

- คลัสเตอร์คอมพิวเตอร์ (Computer Clusters)

  คลัสเตอร์คอมพิวเตอร์ หมายถึงระบบที่ประกอบด้วยคอมพิวเตอร์หลายเครื่องหรือหน่วยประมวลผล (nodes) ทำงานร่วมกันในลักษณะเครือข่าย เพื่อเพิ่มประสิทธิภาพและความสามารถในการประมวลผล

ข้อมูลขนาดใหญ่หรือซับซ้อน โดยคลัสเตอร์คอมพิวเตอร์สามารถแบ่งงานที่ซับซ้อนออกเป็นงานย่อย ๆ และให้แต่ละเครื่องประมวลผลพร้อมกันในลักษณะการประมวลผลแบบขนาน (Parallel Processing) ซึ่งช่วยให้การประมวลผลทำได้รวดเร็วขึ้น อีกทั้งยังมีความสามารถในการขยายตัว (Scalability) นั่นคือการเพิ่มจำนวนเครื่องหรือหน่วยประมวลผลเพิ่มเติมได้โดยไม่ต้องหยุดระบบ ทำให้สามารถเพิ่มประสิทธิภาพการประมวลผลตามความต้องการของผู้ใช้งานได้ นอกจากนี้การใช้คลัสเตอร์คอมพิวเตอร์ช่วยลดต้นทุนได้มากกว่าการใช้ซูเปอร์คอมพิวเตอร์ โดยประเภทของเทคโนโลยีคอมพิวเตอร์คลัสเตอร์สามารถแบ่งออกได้เป็น 4 ประเภทหลักๆ ดังต่อไปนี้

- High-Performance Computing (HPC) Clusters
  คลัสเตอร์ลักษณะนี้ ถูกออกแบบมาเพื่อรองรับงานประมวลผลที่มีความต้องการสูงในเรื่องของความเร็วและพลังประมวลผล โดยใช้แนวคิดใช้การประมวลผลแบบขนาน (Parallel Computing) เพื่อกระจายภาระงานไปยังหลาย ๆ เครื่องในคลัสเตอร์พร้อมกัน ซึ่งจะช่วยเพิ่มประสิทธิภาพและลดเวลาในการประมวลผล
- High-Availability (HA) Clusters
  คลัสเตอร์ลักษณะนี้ ถูกออกแบบมาเพื่อเพิ่มความน่าเชื่อถือและความต่อเนื่องในการทำงานของระบบ หากเครื่องหนึ่งเครื่องใดล้มเหลว เครื่องอื่นๆจะรับหน้าที่แทนทันที ทำให้ระบบยังคงทำงานได้โดยไม่หยุดชะงัก โดยจุดเด่นคือมีการตั้งค่าระบบสำรอง (Failover) หากเกิดความผิดพลาดในเครื่องหลัก การประมวลผลจะถูกย้ายไปยังเครื่องสำรองทันที
- Load-Balancing Clusters
  คลัสเตอร์ลักษณะนี้ ถูกออกแบบมาเพื่อกระจายภาระงานไปยังเครื่องหลาย ๆ เครื่อง เพื่อให้การประมวลผลทำงานอย่างมีประสิทธิภาพและไม่เกิดปัญหาคอขวด (Bottleneck) อันเกิดจากปัญหาเครือข่ายที่ทำให้ภาระงานตกอยู่กับเครื่อง CPU เครื่องใดเครื่องหนึ่งมากเกินไป โดยระบบจะตรวจสอบภาระงานของเครื่องต่าง ๆ ในคลัสเตอร์ หากเครื่องใดมีภาระงานสูง ระบบจะกระจายงานใหม่ไปยังเครื่องที่มีภาระงานน้อยกว่า เพื่อให้การทำงานเป็นไปอย่างสมดุล
- AI Clusters
  คลัสเตอร์ลักษณะนี้ ถูกออกแบบมาเพื่อรองรับการประมวลผลของงานที่เกี่ยวข้องกับปัญญาประดิษฐ์ (Artificial Intelligence) โดยเฉพาะ เช่น การฝึกสอนโมเดล Machine Learning หรือ Deep Learning เป็นต้น โดยมักใช้ GPU และ TPU (Tensor Processing Unit) ที่มีประสิทธิภาพสูงในการประมวลผลงานที่ต้องการความเร็วในการคำนวณสูง รวมถึงการทำงานแบบขนานเพื่อฝึกโมเดล AI ขนาดใหญ่

โครงสร้างพื้นฐานทางด้านเครือข่ายมีหน้าที่จำเพาะคือ ทำให้เกิดการส่งผ่านและเข้าถึงข้อมูลได้จากระยะไกล โดยซูเปอร์คอมพิวเตอร์รุ่นปัจจุบันนั้นโดยส่วนใหญ่ จะถูกสร้างขึ้นบนพื้นฐานแนวคิดของคลัสเตอร์คอมพิวเตอร์และประมวลผลด้วย หน่วยประมวลผลกราฟิก (Graphics Processing Unit: GPU) โดยการประมวลผลร่วมกันกับเครือข่ายความเร็วสูง (High-speed Network) และหน่วยจัดเก็บข้อมูล

โดยในปัจจุบันได้มีการจัดอันดับ TOP500 ที่ทำการจัดอันดับความสามารถของระบบ HPC ทั่วโลก ในด้านประสิทธิภาพในการแก้สมการสมการเชิงเส้นที่มีสัมประสิทธิ์ส่วนใหญ่ไม่เป็นศูนย์ (Dense Linear Equation) โดยอันดับ 1 ของโลกคือ Frontier System ที่สูนย์การวิจัย Oak Ridge National Laboratory ประเทศสหรัญอเมริกา โดยมีประสิทธิภาพอยู่ที่ 1,102 PFlop/s (หมายถึงความสามารถในการดำเนินการทางคณิตศาสตร์แบบทศนิยมจำนวน 1,102 พันล้านล้านครั้งต่อวินาที หรือ 1.102×10^{15} ครั้งต่อวินาที) โดยที่หน่วย Flop/s (Floating Point Operations Per Second) เป็นหน่วยที่ใช้วัดความเร็วของการประมวลผลคอมพิวเตอร์ในงานคำนวณที่มีการใช้จุดทศนิยม

ส่วนในประเทศไทย รัฐบาลได้มีการจัดตั้ง ”ศูนย์ทรัพยากรคอมพิวเตอร์เพื่อการคำนวณขั้นสูง” (NSTDA Supercomputer Center: ThaiSC) โดย LANTA เป็นระบบ HPC ที่ใหญ่ที่สุดในอาเซียนตั้งอยู่ที่ ThaiSC ข้อมูลจากเดือนพฤศจิยายน 2022 LANTA ได้รับการจัดอันดับให้เป็นระบบอันดับที่ 70 ของโลกด้วยการประมวลผลที่ 8.15 Pflop/s

2 วัตถุประสงค์ของการศึกษา (Purpose of the Study)

จากการงานวิจัย A Survey of High-Performance Computing (HPC) Infrastructure in Thailand สามารถสรุปวัตถุประสงค์ของงานวิจัยออกมาเป็นหัวข้อได้ ดังนี้

1. เพื่อสำรวจโครงสร้างพื้นฐานของ HPC ในประเทศไทยตั้งแต่อดีตจนถึงปัจจุบัน
    - ศึกษาถึงโครงสร้างและขนาดของระบบ HPC ที่ใช้ในภาครัฐและเอกชน
    - ศึกษาข้อมูลเกี่ยวกับการจัดซื้อจัดจ้างของรัฐบาล ด้านทรัยากร HPC ในช่วงระยะเวลา 5 ปีที่ผ่านมา
2. เพื่อรวบรวมข้อมูลการใช้งานระบบ HPC ที่มีอยู่ในหน่วยงานต่างๆ ทั้งในภาครัฐและเอกชน
    - ศึกษาและวิเคราะห์จำนวนคอร์ประมวลผล (Cores) และหน่วยจัดเก็บข้อมูล (Storage) ที่มีอยู่ในศูนย์ HPC ทั่วประเทศไทย
    - ศึกษาและสำรวจการใช้งาน HPC ในด้านต่างๆในแต่ละสถาบันหรือหน่วยงาน ไม่ว่าจะเป็นเพื่อการถ่ายโอนข้อมูลหรือการประยุกต์ใช้งานเพื่อรองรับ Artificial Intellenge, Machine Learning, การประมวลผลภาษาธรรมชาติ (NLP), และ TensorFlow เป็นต้น ไปจนถึงการประยุกต์ใช้ทางวิทยาศาสตร์เชิงชีวภาพ
3. เพื่อศึกษาความท้าทายและปัญหาที่เกี่ยวข้องกับการใช้งานระบบ HPC ในประเทศไทย
    - ศึกษาในประเด็นเกี่ยวกับการถ่ายโอนข้อมูลขนาดใหญ่ (Transfering of Large Data) และข้อจำกัดทางด้านเครือข่าย คือข้อกำจัดของ Bandwith และความล่าช้าในการถ่ายโอนข้อมูลที่เกิดจากปัญหาคุณภาพเครือข่ายที่ไม่สม่ำเสมอ อันเป็นอุปสรรคต่อการถ่ายโอนข้อมูลระหว่างศูนย์เครือข่าย เนื่องมาจากโครงสร้างพื้นฐานที่รองรับ HPC ยังไม่สมบูรณ์ ส่งผลให้การใช้ HPC ในการประมวลผลข้อมูลจากหลากหลายแหล่งเป็นไปได้ยาก
    - ศึกษาอุปสรรคในการพัฒนาโครงสร้างพื้นฐานของ HPC ซึ่งเป็นผลมาจาก 3 ปัจจัยด้วยกันคือ ปัญหาการเพิ่มขอบเขตการประยุกต์ใช้งานให้กับ HPC (Application Field), ปัญหาการปรับตัวของ HPC ไปสู่การใช้งานบนคลาวด์ของภาครัฐ และปัญหาการขาดแคลนผู้เชี่ยวชาญที่มีความรู้ทางด้าน HPC ที่ยังมีไม่มากพอในประเทศ
4. เพื่อสำรวจผลกระทบและประโยชน์ที่ HPC มอบให้ภาคอุตสาหกรรมและงานวิจัยในประเทศไทย
    - ศึกษาการประยุกต์ใช้งาน HPC ในการสนับสนุนการวิจัยในด้านต่างๆ เช่น การคำนวณทางฟิสิกส์และการทำนายสภาพอากาศ เป็นต้น
5. เพื่อเสนอแนะแนวทางการพัฒนาการวิจัยด้าน HPC ในอนาคต

- ศึกษาแนวทางในการพัฒนาและขยายโครงสร้างพื้นฐานของ HPC ในประเทศไทย
- เสนอแนะนโยบาย แนวทางการพัฒนาเทคโนโลยีใหม่ และการจัดสรรทรัพยากรในการพัฒนาระบบ HPC

3 วิธีการศึกษา (Methodology)

วิธีการศึกษาในแบบสำรวจนี้ แบ่งออกเป็นขั้นตอนหลักๆ ดังนี้

1. การรวบรวมข้อมูลจากหน่วยงานที่ใช้งาน HPC

   ผู้วิจัยได้รวบรวมข้อมูลเกี่ยวกับระบบ HPC ที่ใช้งานในประเทศไทยจากหน่วยงานต่างๆ ทั้งภาครัฐและเอกชน รวมถึงมหาวิทยาลัย โดยใช้วิธีการเก็บข้อมูลผ่านการจัดซื้อจัดจ้างของรัฐบาลในช่วง 5 ปีที่ผ่านมา นอกจากนี้ยังรวบรวมข้อมูลจากแหล่งข้อมูลอื่นๆ เช่น ระบบ HPC ที่มีอยู่ในประเทศไทย การใช้ซอฟต์แวร์ และการพัฒนาระบบเครือข่าย

2. การรวบรวมข้อมูลของระบบ HPC ในประเทศไทย

   ข้อมูลที่ได้รวบรวมมานั้นถูกนำมาวิเคราะห์เพื่อประเมินลักษณะต่างๆ ของระบบ HPC เช่น จำนวนคอร์ประมวลผล หน่วยความจำ พื้นที่จัดเก็บข้อมูล ซอฟต์แวร์ที่ใช้งาน และการเชื่อมต่อเครือข่าย การวิเคราะห์นี้ช่วยให้เห็นภาพรวมของโครงสร้างพื้นฐานของ HPC ในประเทศไทยโดยรวม

3. การศึกษาเกี่ยวกับการถ่ายโอนข้อมูล

   มีการศึกษาเกี่ยวกับปัญหาการถ่ายโอนข้อมูลขนาดใหญ่ระหว่างศูนย์ HPC โดยเน้นไปที่ความท้าทายที่เกี่ยวข้องกับการใช้เครือข่าย UniNet ซึ่งเป็นเครือข่ายการศึกษาและวิจัยของประเทศไทย ข้อมูลที่ได้จากการศึกษานี้ถูกนำมาวิเคราะห์เพื่อหาวิธีการพัฒนาการเชื่อมต่อและการส่งข้อมูลที่มีประสิทธิภาพมากขึ้น

4. การสัมภาษณ์และสอบถามจากผู้เชี่ยวชาญทางด้าน HPC

   นอกจากการรวบรวมข้อมูลทางกายภาพ ผู้วิจัยยังได้สัมภาษณ์และเก็บแบบสอบถามจากบุคคลในวงการ HPC ในประเทศไทย เพื่อให้ได้มุมมองเกี่ยวกับอุปสรรค ปัญหา และความท้าทายในการใช้งาน HPC ข้อมูลนี้ถูกนำมาใช้ในการเสนอแนะวิธีการพัฒนาระบบในอนาคต

5. การประเมินและเสนอแนะ

   หลังจากรวบรวมและวิเคราะห์ข้อมูลต่างๆ ผู้วิจัยได้ประเมินผลและนำเสนอข้อเสนอแนะเพื่อพัฒนาระบบ HPC ในประเทศไทย โดยเน้นไปที่การพัฒนาระบบให้รองรับการประมวลผลที่มีขนาดใหญ่ขึ้น และการแก้ไขปัญหาการถ่ายโอนข้อมูลที่อาจเกิดขึ้น

จากวิธีการศึกษาดังกล่าว ผู้วิจัยสามารถให้ข้อมูลเชิงลึกเกี่ยวกับสถานะของระบบ HPC ในประเทศไทย พร้อมกับเสนอแนวทางในการปรับปรุงและพัฒนาในอนาคต

4 ผลการศึกษา (Findings)

การค้นพบสำคัญในเปเปอร์นี้สามารถแบ่งออกเป็นหัวข้อหลักๆ ดังนี้

1. โครงสร้างและความเป็นมาของระบบ HPC ในประเทศไทย

    - ผู้วิจัยพบว่ามีการลงทุนและการพัฒนาระบบ HPC (High-Performance Computing) อย่างต่อเนื่องในประเทศไทย ทั้งในหน่วยงานภาครัฐและเอกชน เช่น มหาวิทยาลัย หน่วยงานวิจัย และองค์กรของรัฐบาล โดยในช่วง 5 ปีที่ผ่านมา มีการจัดซื้อจัดจ้างอุปกรณ์และระบบ HPC เป็นจำนวนมาก ทำให้ระบบ HPC ในประเทศไทยมีการพัฒนาอย่างรวดเร็ว
    - เหตุการณ์สำคัญที่เกิดขึ้น ในปี 1980s วิทยาศาสตร์การคำนวณได้กลายเป็นสาขาการวิจัยใหม่ในประเทศไทย หน่วยงาน ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (เนคเทค สวทช. หรือ NECTEC) ได้จัดตั้งศูนย์ HPCC ในประเทศ โดยได้ทำการวิจัยโดยการใช้เครื่องคอมพิวเตอร์ที่มีความสามารถสูง (High-performance workstations) ในช่วงแรกๆ นักวิจัยไทยใช้เครื่อง workstations เพื่อทำการคำนวณเชิงซับซ้อน เช่น DEC Alpha, HP Apollo, IBM RS6000 และ SUN SPARC เป็นต้น
    - เหตุการณ์สำคัญที่เกิดขึ้น ในปี 1990s เช่น การได้มาของ CRAY EL98 และ SGI Power Challenge XL เป็นต้น อีกทั้งมหาวิทยาลัยเกษตรศาสตร์ (KU) ,มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี (KMUTT) ได้รับ IBM SP2 systems และสถาบันเทคโนโนลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (KMITL) ได้รับ 8-processor convex system
    - เหตุการณ์สำคัญที่เกิดขึ้น ในปี 1995-1997 ได้มีการพัฒนาแนวคิด Beowulf cluster ขึ้น จนต่อมาในปีในปี 1999 ที่มหาวิทยาลัยเกษตรศาสตร์ (KU) ซึ่งเป็นหนึ่งในสถาบันที่ร่วมทำการพัฒนาคลัสเตอร์ขนาดใหญ่ โดยใช้เครื่องคอมพิวเตอร์จำนวน 72 เครื่องที่เรียกว่า PIRUN Cluster ซึ่งพัฒนาร่วมกับการใช้ซอฟต์แวร์จัดการคลัสเตอร์ เช่น SMILE Cluster Management System และการใช้เครื่องมือการจัดการงานแบบ batch scheduler เช่น SQMS (Simple Queue Management System) ทำให้สามารถแบ่งทรัพยากรการประมวลผลและจัดการงานที่ส่งเข้าไปในระบบคลัสเตอร์ได้อย่างมีประสิทธิภาพ
    - เหตุการณ์สำคัญที่เกิดขึ้น ในปี 2001s หลังจากมีการพัฒนา cluster ขนาดใหญ่เช่น การสร้าง Beowulf cluster และการขยายการใช้ grid computing เป็นอีกหนึ่งความก้าวหน้าในด้าน HPC ซึ่งโครงการ ThaiGrid ได้เริ่มขึ้นในปีนี้ โดยโครงการนี้ความร่วมมือระหว่างมหาวิทยาลัยเกษตรศาสตร์ (KU) และมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี (KMUTNB) โดยในต่อมาก็มีมหาลัยอื่นๆเข้าร่วมมากขึ้น โครงการนี้เป็นก้าวแรกของการสร้างโครงสร้างพื้นฐานการประมวลผลแบบ Grid ในประเทศไทย โดยนำเสนอโอกาสในการใช้ HPC ในการวิจัยด้านต่างๆ ที่มีความสำคัญระดับประเทศ เช่น การพยากรณ์อากาศ การวิเคราะห์ข้อมูลขนาดใหญ่ และงานวิจัยด้านวิทยาศาสตร์อื่นๆ เป็นต้น

- ในช่วงปี 2006–2008 โครงการ ThaiGrid ได้รับการสนับสนุนจากกระทรวงวิทยาศาสตร์และเทคโนโลยี มีการจัดตั้งศูนย์กลางระดับชาติที่ชื่อว่า Thai National Grid Center (TNGC) โดยศูนย์นี้มีบทบาทในการบริหารจัดการโครงสร้างพื้นฐานและทรัพยากรของ Grid สำหรับการวิจัยและการศึกษา รวมถึงการสนับสนุนโครงการด้านวิทยาศาสตร์และวิศวกรรมในประเทศ
- แม้ว่าโครงการ TNGC จะประสบความสำเร็จในช่วงเริ่มต้น แต่ก็ต้องหยุดดำเนินการในปี 2008 เนื่องจากปัญหาด้านงบประมาณ อย่างไรก็ตาม บทเรียนที่ได้รับจากโครงการ TNGC ถือเป็นการวางรากฐานสำคัญให้กับประเทศไทยในการพัฒนา HPC ในอนาคต โดยเฉพาะการสนับสนุนการพัฒนาโครงสร้างพื้นฐานในภาคการศึกษาและการวิจัย
- ในปี 2011 มีการจัดตั้ง National e-Science Infrastructure Consortium มีการก่อตั้งขึ้นเพื่อเป็นศูนย์กลางในการสนับสนุนและแบ่งปันทรัพยากร HPC ระหว่างมหาวิทยาลัยและสถาบันวิจัยต่างๆ ในประเทศ โดยสมาชิกของคณะกรรมการนี้ประกอบด้วยมหาวิทยาลัยและสถาบันวิจัยหลายแห่ง โดยมีเป้าหมายในการสนับสนุนการวิจัยทางวิทยาศาสตร์ที่ต้องการใช้ HPC ในการประมวลผลข้อมูลขนาดใหญ่
- ใบปัจจุบันทรัพยากร HPC ถูกกระจายอยู่ใน 13 มหาวิทยาลัย, 4 สถาบันวิจัย, 3 หน่วยงานของรัฐ, 1 รัฐวิสาหกิจ (การไฟฟ้าฝ่ายผลิตแห่งประเทศไทย หรือ EGAT) และ 1 บริษัทเอกชน (Siam Cement Group หรือ SCG) ซึ่งครอบคลุมหลายภาคส่วน องค์กรเหล่านี้ดำเนินการระบบ HPC ที่มีประสิทธิภาพเพื่อรองรับการประมวลผลที่หลากหลาย โดยหน่วยงานรัฐบาลที่สำคัญในการดูแลโครงสร้างพื้นฐาน HPC ได้แก่ กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม (MHESI) และกระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม (MDES) สองกระทรวงนี้มีบทบาทสำคัญในการดูแลและดำเนินการระบบ HPC หลักของประเทศ
- การแบ่งประเภทของ HPC ในประเทศไทย จากข้อมูลที่รวบรวมมา ระบบ HPC ในประเทศไทยสามารถแบ่งออกเป็นสองประเภทหลักคือ HPC Clusters ที่ใช้หน่วยประมวลผลกราฟิก (GPU) และ NVIDIA DGX เซิฟเวอร์ โดย HPC Clusters ที่ใช้ GPUs ในประเทศไทยสามารถพบได้ในหลายสถาบัน เช่น NSTDA Supercomputer Center (ThaiSC) ซึ่งใช้หน่วยประมวลผลแบบ NVIDIA A100 และ V100 GPUs ในการประมวลผลขนานเพื่อรองรับการวิจัยที่ต้องใช้พลังการประมวลผลสูง และ ในประเทศไทยมีการใช้ DGX Servers สำหรับงานที่เกี่ยวข้องกับ AI และการจำลองโมเดลวิทยาศาสตร์ ตัวอย่างเช่น LANTA และ NECTEC ใช้ NVIDIA DGX Servers เพื่อสนับสนุนการวิจัยทาง AI และการประมวลผลทางวิทยาศาสตร์ เป็นต้น
- ศูนย์คอมพิวเตอร์ระดับสูงที่สำคัญในประเทศไหนมี LANTA ที่ดำเนินการโดย ThaiSC มีจำนวนการประมวลผลสูงที่สุดในประเทศไทยด้วยจำนวน 33,024 คอร์ และประสิทธิภาพการประมวลผลที่ 8.15 PFlops/s ซึ่งถือว่าเป็นทรัพยากร HPC ที่มีความสำคัญในประเทศ
- องค์กรทีใช้ทรัพยากร HPC ในประเทศไทย เช่น CMKL University (ความร่วมมือระหว่าง Carnegie Mellon University และ King Mongkut’s Institute of Technology Ladkrabang (KMITL)) ทั้งยังเป็นศูนย์กลางด้านการวิจัยและพัฒนาด้าน HPC ในประเทศไทย มีทรัพยากร

HPC ซึ่งก็คือ APEX เป็นแพลตฟอร์มการประมวลผลแบบประสิทธิภาพสูง (High-Performance Computing หรือ HPC) และโครงสร้างพื้นฐานสำหรับการจัดเก็บข้อมูล ที่สามารถสนับสนุนการวิจัยและพัฒนาในหลายสาขา เช่น AI, Machine Learning, และ Data Science เป็นต้น โดยมีลักษณะของทรัพยากรคือ จำนวนโหนด 6 โหนด มีทั้งหมด 768 cores หน่วยประมวลผลกราฟิก (GPUs) ใช้การ์ด NVIDIA A100 จำนวน 48 หน่วย และการ์ด Intel Phi 1 หน่วย หน่วยความจำสำหรับแต่ละโหนดคือ 1 TB มีหน่วยความจำรวมของระบบคือ 6 TB ระบบมีพื้นที่จัดเก็บข้อมูลทั้งหมด 300 TB มีการเชื่อมต่อ (Interconnections) ใช้การเชื่อมต่อที่ความเร็ว 200 Gbps InfiniBand

- จุดประสงค์ของ APEX ที่ CMKL คือการสร้างโครงสร้างพื้นฐานที่มีความยืดหยุ่น สามารถรองรับการใช้งานของผู้ใช้จำนวนมากในเวลาเดียวกัน โดยไม่ลดทอนประสิทธิภาพการทำงาน ซึ่งทำให้ APEX เป็นหนึ่งในแพลตฟอร์ม HPC ที่สำคัญสำหรับนักวิจัยทั้งในระดับประเทศและระดับนานาชาติ
- สำหรับระบบปฏิบัติการและการประยุกต์ใช้งานที่ใช้ในปี 2022 ระบบ HPC ของ CMKL เริ่มใช้งานตั้งแต่ปี 2020 ใช้ระบบปฏิบัติการ Ubuntu ซึ่งเป็นระบบปฏิบัติการ Linux ที่มีความนิยมสูง สำหรับงานวิจัยด้าน HPC ใช้ประมวลผลด้าน AI, Machine Learning (ML), และ TensorFlow นอกจากนี้ยังมีการใช้งานด้าน Natural Language Processing (NLP) และการวิเคราะห์เชิงลึกในด้าน AI for Health ซึ่งรวมถึงการประยุกต์ AI เพื่อการท่องเที่ยว (AI for Tourism)

2. การถ่ายโอนข้อมูลขนาดใหญ่สำหรับการประมวลผลประสิทธิภาพสูง (HPC) การถ่ายโอนข้อมูลระหว่างหน่วยงานหรือสิ่งอำนวยความสะดวกต่าง ๆ ซึ่งส่วนใหญ่จะเกี่ยวข้องกับข้อมูลทางวิทยาศาสตร์ เช่น อุปกรณ์การวัดทางวิทยาศาสตร์ (Science instruments) อุปกรณ์ IoT และโซเชียลมีเดีย โดยข้อมูลเหล่านี้เพิ่มขึ้นในปริมาณอย่างรวดเร็วในช่วงไม่กี่ปีที่ผ่านมา จำเป็นต้องมีระบบเครือข่ายที่มีความเร็วสูงเพื่อรองรับการประมวลผล การรวบรวม และการแสดงผลข้อมูลเหล่านี้

    - การเพิ่มขึ้นของข้อมูลขนาดใหญ่เป็นปัจจัยสำคัญในงานประมวลผลประสิทธิภาพสูง (HPC) เนื่องจากแหล่งข้อมูลที่หลากหลาย เช่น อุปกรณ์วิทยาศาสตร์ อุปกรณ์ IoT และโซเชียลมีเดีย กำลังสร้างชุดข้อมูลจำนวนมหาศาลซึ่งเติบโตอย่างรวดเร็ว ในประเทศไทย ข้อมูลจากหน่วยงานวิจัยหลายแห่ง เช่น SLRI (Synchrotron Light Research Institute), TMD (Thai Meteorological Department) และ NBT (National Biobank of Thailand) มีบทบาทสำคัญในกระบวนการเก็บรวบรวมและประมวลผลข้อมูลขนาดใหญ่ ชุดข้อมูลเหล่านี้ต้องอาศัยการเชื่อมต่อเครือข่ายความเร็วสูงสำหรับการรวบรวม การประมวลผล และการแสดงผล เพื่อให้สามารถใช้ข้อมูลเหล่านี้ในงานวิจัยทางวิทยาศาสตร์และการพัฒนาทางเทคโนโลยีต่าง ๆ ได้อย่างมีประสิทธิภาพ
    - REN Programs (Research and Education Network) เป็นโครงการที่ขับเคลื่อนโดยความร่วมมือของรัฐบาลและองค์กรต่างๆ ทั้งในระดับประเทศและนานาชาติ เพื่อสนับสนุนการเชื่อม

ต่อเครือข่ายความเร็วสูงสำหรับการศึกษาและการวิจัย โครงการเหล่านี้มีบทบาทสำคัญในการพัฒนาการส่งผ่านข้อมูลขนาดใหญ่และการทำงานร่วมกันระหว่างสถาบันการศึกษาและศูนย์วิจัยในหลายประเทศ โดยในประเทศไทย โครงการ REN ขับเคลื่อนโดยการสนับสนุนจากรัฐบาลไทย รวมถึงโครงการเครือข่ายอย่าง Internet2, ESnet, DFN, JGN, MYREN, APAN, GE ANT, และ TEIN ที่ช่วยสร้างโครงข่ายเชื่อมต่อที่มีประสิทธิภาพสำหรับการประมวลผลข้อมูลขนาดใหญ่ และการสนับสนุนโครงสร้างพื้นฐานทางวิทยาศาสตร์และการศึกษา

- เครือข่ายหลักสำหรับการศึกษาและวิจัย (UniNet) เป็นเครือข่ายที่ให้บริการด้านการศึกษาวิจัยระดับสูงในประเทศไทย โดยเริ่มดำเนินการตั้งแต่ปี 1996 โดยมีจุดมุ่งหมายเพื่อสนับสนุนสถาบันการศึกษาและองค์กรวิจัยทั่วประเทศ โครงข่ายหลักของ UniNet ได้พัฒนาความสามารถของแบนด์วิดท์ตั้งแต่ T1, OC3, OC12, OC48 จนถึง 10 Gbps Ethernet และล่าสุดได้ขยายเป็น 100 Gbps Ethernet ทำให้สามารถเชื่อมต่อกับศูนย์การศึกษาและวิจัยในระดับสูงได้อย่างมีประสิทธิภาพ โดยมีความสามารถในการรองรับการถ่ายโอนข้อมูลขนาดใหญ่ระหว่างจังหวัดและพื้นที่การศึกษา อีกทั้งมีการติดตั้งลิงก์ 10 Gbps หลายจุดเพื่อเชื่อมโยงศูนย์ HPC ต่างๆ ในเครือข่ายเดียวกัน
- ปัญหาคอขวดในการถ่ายโอนข้อมูล (Bottlenecks in Data Transfer) เกิดขึ้นจากข้อจำกัดในโครงสร้างพื้นฐานของเครือข่ายและอุปกรณ์รักษาความปลอดภัยที่ใช้ในการรับส่งข้อมูล โดยเฉพาะโมเดลบริการอินเทอร์เน็ตแบบ best-effort service ซึ่งสร้างความล่าช้าในกระบวนการส่งข้อมูลระหว่างจุดปลายทางสองจุด ระบบเครือข่ายมักประสบปัญหาการติดขัดเนื่องจากอุปกรณ์รักษาความปลอดภัย เช่น ไฟร์วอลล์หรือระบบรักษาความปลอดภัยอื่น ๆ ที่ทำให้ความเร็วในการส่งข้อมูลลดลง เทคนิคบางอย่างถูกเสนอมาเพื่อบรรเทาปัญหานี้ เช่น การใช้เครือข่าย overlay เครือข่ายที่มีความเฉพาะเจาะจง (dedicated links) และการจัดการคุณภาพการให้บริการ (QoS) เพื่อเพิ่มความรวดเร็วและประสิทธิภาพในการรับส่งข้อมูล
- การพัฒนาเครือข่ายแบบใหม่ (New Network Development) มุ่งเน้นไปที่การสร้างโครงข่ายที่มีความสามารถในการปรับตัวสูงและมีประสิทธิภาพในการถ่ายโอนข้อมูลระหว่างศูนย์ HPC (High-Performance Computing) เพื่อให้ตรงตามความต้องการที่แตกต่างกันของผู้ใช้งาน โดยเครือข่ายแบบใหม่ เช่น เครือข่ายที่กำหนดได้ด้วยซอฟต์แวร์ (Software-Defined Networks: SDN) เป็นโครงข่าย overlay ได้ถูกนำมาใช้เพื่อเพิ่มการควบคุมประสิทธิภาพและความปลอดภัยในการถ่ายโอนข้อมูล ซึ่งทำให้ผู้ดูแลระบบสามารถปรับการทำงานของเครือข่ายได้อย่างยืดหยุ่นตามความต้องการเฉพาะของแต่ละโครงการ โดยเฉพาะการถ่ายโอนข้อมูลที่มีขนาดใหญ่และซับซ้อน อีกทั้งยังมี Demilitarized Zone (DMZ) เป็นแนวคิดที่ถูกนำมาใช้ในโครงการพัฒนาเครือข่ายใหม่ของ UniNet เพื่อให้เกิดการแยกเครือข่ายที่มีความปลอดภัยสูงระหว่างศูนย์การศึกษาวิจัยต่างๆ โดย DMZ ช่วยลดความเสี่ยงด้านความปลอดภัยและช่วยควบคุมการถ่ายโอนข้อมูลระหว่างปลายทางสองจุด โดยเป็นเสมือนพื้นที่กักกันที่ปลอดภัยสำหรับข้อมูลสำคัญ ซึ่งจะสร้างเกราะป้องกันเครือข่ายวิจัยไม่ให้ถูกโจมตีจากภายนอก

- การถ่ายโอนข้อมูลที่มีความเร็วสูง (High-speed Data Transfer) การถ่ายโอนข้อมูลที่มีความเร็วสูง (High-speed Data Transfer) นั้นเป็นผลจากการใช้เครือข่าย UniNet ที่พัฒนาขึ้นเพื่อการถ่ายโอนข้อมูลขนาดใหญ่ระหว่างศูนย์วิจัยและการศึกษา โดยใช้ความเร็วสูงถึง 7.2 Gbps ผ่านระบบเชื่อมต่อแบบหลายสตรีม (multi-stream parallel data transfer) ซึ่งแสดงให้เห็นประสิทธิภาพในการถ่ายโอนข้อมูลปริมาณมากในระยะเวลาอันสั้น ผลลัพธ์จากการถ่ายโอนข้อมูลนี้สามารถลดระยะเวลาการถ่ายโอนข้อมูลลงได้อย่างมีนัยสำคัญ เช่น การถ่ายโอนข้อมูลขนาด 5 TB จะใช้เวลาเพียง 2.5 ชั่วโมง ซึ่งแสดงให้เห็นถึงศักยภาพในการสนับสนุนงานวิจัยที่ต้องการข้อมูลขนาดใหญ่

3. ความท้าทายและโอกาสในอนาคตของการพัฒนา HPC

- ความต้องการทรัพยากร (Creating Needs) ในประเทศไทย ทรัพยากร HPC ถูกแบ่งออกเป็น 3 ประเภทหลัก ได้แก่ Small Clusters ในมหาวิทยาลัยที่สนับสนุนการวิจัยและการศึกษา, Application-Specific Resources สำหรับงานเฉพาะทาง เช่น การจำลองวิศวกรรม, และ Large-Scale HPC Facilities ที่ใช้สำหรับการประมวลผลที่ซับซ้อน เพื่อรองรับงานวิจัยที่ต้องการทรัพยากรสูง เป็นต้น โดยนักวิจัยสามารถใช้ทรัพยากรจาก ThaiSC หรือบริการ Cloud-based public service ได้ การพัฒนาและรักษาโครงสร้างพื้นฐาน HPC เป็นเรื่องสำคัญที่ต้องมีการประหยัดและการพัฒนาเทคโนโลยีในสังคม การพยากรณ์อากาศ การจำลองทางวิทยาศาสตร์ เช่น ชีววิทยาและเคมี ถือเป็นส่วนหนึ่งของงานที่ HPC รองรับ ThaiSC เป็นศูนย์กลาง HPC ในประเทศไทยมาตั้งแต่ปี 1992 โดยมีบทบาทสำคัญในการสนับสนุนงานวิจัยใหม่ๆ เช่น การพยากรณ์ฝุ่น PM2.5 และการใช้ข้อมูลจีโนมในการวินิจฉัย เป็นต้น เพื่อให้โครงสร้างพื้นฐาน HPC สามารถดำเนินการได้อย่างยั่งยืน การประหยัดทรัพยากรและการพัฒนาเทคโนโลยีต้องได้รับการสนับสนุนจากสังคม โดยเฉพาะในด้านการพยากรณ์อากาศ, การจำลองทางชีววิทยา, เคมี, และวิทยาศาสตร์อื่นๆ ดังที่ได้เคยกล่าวไว้ข้างต้น
- การดำเนินงานของโครงสร้างพื้นฐาน (Operating Infrastructure) คือปัญหาที่สองในการจัดการโครงสร้างพื้นฐาน HPC โดยเน้นที่ความผันผวนของความต้องการใช้งาน HPC ซึ่งส่งผลให้เกิดการใช้งานต่ำลง (Low utilization) และระบบที่มีการใช้งานต่ำมักจะได้รับการสนับสนุนด้านงบประมาณที่ลดลง ในขณะที่อายุการใช้งานเฉลี่ยของทรัพยากรคอมพิวเตอร์อยู่ที่ประมาณ 7 ปี นอกจากนี้ การเติบโตของอุตสาหกรรมคลาวด์ยังมีบทบาทสำคัญในการลดช่องว่างระหว่างการใช้งาน HPC แบบดั้งเดิมกับคลาวด์ ขณะนี้มีบริการคลาวด์ที่สามารถนำเสนอ HPC ที่มีประสิทธิภาพเช่นเดียวกับระบบดั้งเดิม โดยคลาวด์มีความยืดหยุ่นมากขึ้น ทั้งในด้านความหลากหลายของบริการ ค่าใช้จ่าย และความสามารถในการปรับขนาด (scalability) เช่น บริการจาก AWS, Microsoft Azure และ Google ที่ให้บริการ HPC ผ่านระบบคลาวด์ เป็นต้น
  การใช้งาน HPC ผ่านคลาวด์กลายเป็นสิ่งที่น่าสนใจสำหรับภาคเอกชนและสถาบันวิจัยหลายแห่ง โดยเฉพาะในกลุ่มที่กังวลเรื่องค่าใช้จ่ายและความคุ้มค่าจากการลงทุนในระบบ HPC ภายใน

(On-premise) ข้อได้เปรียบที่เห็นได้ชัดคือคลาวด์สามารถขยายตัวตามความต้องการได้ง่ายและมีความยืดหยุ่นมากกว่าระบบภายใน

อย่างไรก็ตามสำหรับองค์กรภาครัฐ การจัดซื้อจัดจ้างที่เกี่ยวข้องกับ HPC ยังไม่ได้ถูกปรับให้ใช้รูปแบบการเช่าซื้อคลาวด์เช่นเดียวกับการจัดซื้อสาธารณูปโภคอื่นๆ เนื่องจากยังมีความท้าทายในการดำเนินการทางเอกสารของภาครัฐที่ขาดความยืดหยุ่น แต่เริ่มมีการพัฒนาโมเดลการจัดซื้อที่ยืดหยุ่นมากขึ้นเพื่อรองรับการใช้งานคลาวด์

สรุปโดยภาพรวมแล้วแนวทางปัญหาด้านการดำเนินงานอาจมีทางออกคือ การพัฒนาเทคโนโลยีเครือข่ายและระบบเสมือนยังมีความสำคัญในการเพิ่มประสิทธิภาพการถ่ายโอนข้อมูลขนาดใหญ่ และการสร้างกลไกใหม่ๆ เช่น Data Transfer Nodes (DTNs), SDWAN และ SDN จะช่วยให้ HPC มีประสิทธิภาพในการถ่ายโอนข้อมูลขนาดใหญ่ได้ดียิ่งขึ้น ซึ่งเทคโนโลยีเหล่านี้เป็นสิ่งที่ชุมชน HPC ในปัจจุบันต้องนำไปใช้

- การพัฒนาทรัพยากรมนุษย์ (Human Resource Development) มีมหาวิทยาลัยจำนวนน้อยที่เปิดสอนหลักสูตรเกี่ยวกับ HPC เช่น การประมวลผลแบบขนาน (Parallel Computing) การประมวลผลแบบกระจาย (Distributed Computing) และ Big Data เป็นต้น โดยมีนักวิจัยด้านนี้ประมาณ 50 คน มีนักศึกษาที่ศึกษา Big Data และ Data Science ประมาณ 150 คน และมีนักศึกษาที่ศึกษา HPC ประมาณ 450 คนต่อปี นอกจากนี้ยังมีนักศึกษาประมาณ 300 คนที่เข้าร่วมเวิร์คช็อป การแข่งขัน และกิจกรรมที่เกี่ยวข้องกับ HPC ในชุมชน HPC ของประเทศไทย

  ในประเทศไทยยังคงต้องการบุคลากรที่ดูแลระบบ HPC มีหน้าที่หลายด้าน เช่น การติดตั้งฮาร์ดแวร์และซอฟต์แวร์ การจัดการระบบไฟล์, การตั้งค่าและดูแลศูนย์ข้อมูล, การจัดการตารางเวลาการประมวลผล, และการแก้ปัญหาคอขวด (Bottleneck) เป็นต้น ในระบบ บุคลากรเหล่านี้ควรมีพื้นฐานที่แข็งแกร่งด้านวิทยาการคอมพิวเตอร์และต้องมีทักษะพิเศษที่เกี่ยวข้องกับ HPC อีกเรื่องหนึ่งคือ

  การพัฒนาทักษะด้านการพัฒนาแอปพลิเคชันและเครื่องมือสำหรับการประมวลผลประสิทธิภาพสูงมีความสำคัญอย่างยิ่ง ซึ่งจะช่วยเพิ่มประสิทธิภาพในการใช้ทรัพยากร HPC นอกจากนี้ กลุ่มที่สามยังครอบคลุมการพัฒนาอินเทอร์เฟซแอปพลิเคชัน (API) และเครื่องมือการประมวลผลบนเว็บเพื่อเพิ่มการเชื่อมต่อระหว่างผู้ใช้ ข้อมูล และบริการต่างๆ

  นอกจากนี้ ผู้เขียนบทความวิจัยนี้ยังเสนอว่า ความสมดุลระหว่างการพัฒนาบุคลากร และการขยายโครงสร้างพื้นฐาน HPC ควรมีการพัฒนาในสองด้านนี้อย่างสมดุล เนื่องจากการเรียนรู้ HPC ต้องใช้เวลาและความพยายามอย่างมากในการทำความเข้าใจและฝึกฝนทักษะต่างๆ

การค้นพบเหล่านี้สะท้อนให้เห็นถึงการพัฒนาของระบบ HPC ในประเทศไทย รวมถึงความท้าทายและโอกาสในอนาคตที่สามารถนำไปสู่การปรับปรุงและพัฒนาระบบให้มีประสิทธิภาพมากยิ่งขึ้น

5 การอภิปรายผล (Discussion)

จากความก้าวหน้าทางเทคโนโลยี HPC ในประเทศไทยที่เราได้มีการกล่าวถึงในหัวข้อข้างต้น ทำให้ทางผู้เขียนตระหนักถึงการต่อยอดและการพัฒนางานวิจัยทางด้าน HPC ในบางหัวข้อที่อาจเป็นการส่งเสริมให้องค์ความรู้ของประเทศไทยในด้าน HPC ได้รับการพัฒนาและเป็นประโยชน์ต่อวงการ HPC ของไทยมากขึ้น ทั้งหมด 7 หัวข้อการวิจัยที่เกี่ยวข้อง ดังนี้

1. การประยุกต์ใช้ HPC ในการฝึกสอนโมเดลปัญญาประดิษฐ์ขนาดใหญ่ (Application of HPC in Training Large-Scale AI and Machine Learning Models)

   การประยุกต์ใช้ HPC (High-Performance Computing) ในการฝึกสอนโมเดล AI และ Machine Learning ขนาดใหญ่นั้นมีความสำคัญอย่างมาก เนื่องจากการฝึกสอนโมเดลที่ซับซ้อน เช่น Deep Learning จำเป็นต้องใช้ทรัพยากรในการประมวลผลสูง การใช้ HPC จะช่วยเร่งกระบวนการฝึกสอนโดยการกระจายงานประมวลผลไปยังคลัสเตอร์คอมพิวเตอร์ที่มีสมรรถนะสูง ทำให้สามารถจัดการข้อมูลปริมาณมากและลดระยะเวลาในการฝึกโมเดล นอกจากนี้ การพัฒนาอัลกอริทึมที่สามารถทำงานแบบขนานบน HPC จะช่วยเพิ่มประสิทธิภาพในการจัดการข้อมูลและประมวลผลโมเดล AI อย่างรวดเร็วและมีประสิทธิภาพมากขึ้น ซึ่งจะเป็นประโยชน์อย่างมากในการวิจัยและพัฒนาด้าน AI และ Machine Learning ในอนาคต

   ตัวอย่างงานวิจัยที่เกี่ยวข้อง มีดังนี้

   - "**Large Scale Distributed Deep Networks**" โดย Jeff Dean, Greg Corrado และ Rajat Monga

     งานวิจัยนี้มีคอนเซปหลักเกี่ยวกับการพัฒนาวิธีการฝึกสอนโมเดล deep learning ที่มีขนาดใหญ่โดยใช้เครือข่ายคอมพิวเตอร์แบบกระจาย (distributed computing) งานวิจัยนี้นำเสนอการพัฒนาเฟรมเวิร์กที่ชื่อว่า DistBelief ที่สามารถใช้คลัสเตอร์คอมพิวเตอร์ที่ประกอบด้วยเครื่องหลายพันเครื่องเพื่อฝึกโมเดลขนาดใหญ่ โดยมีอัลกอริทึมหลักสองตัวคือ Downpour SGD และ Sandblaster ซึ่งช่วยเร่งกระบวนการฝึกสอนโมเดล deep learning ได้อย่างมีประสิทธิภาพ งานวิจัยนี้ประสบความสำเร็จในการฝึกโมเดล Deep network ที่มีขนาดใหญ่ถึง 30 เท่าจากงานวิจัยเดิม และยังได้แสดงให้เห็นถึงความสามารถในการปรับใช้กับงานด้านการรู้จำภาพ (ImageNet) และการรู้จำเสียง งานวิจัยนี้ให้ประโยชน์ในด้านการเพิ่มประสิทธิภาพและความเร็วในการฝึกสอนโมเดล Deep learning ที่สามารถนำไปประยุกต์ใช้กับอัลกอริทึม Machine learning อื่นๆ

   - "**Scaling Distributed Machine Learning with the Parameter Server**" โดย Mu Li, David G. Andersen และ Alexander J. Smola

     งานวิจัยนี้นำเสนอเฟรมเวิร์ก Parameter Server สำหรับการแก้ปัญหาใน Distributed Machine Learning โดยการกระจายข้อมูลและงานไปยัง Worker nodes ในขณะที่ Server

nodes จัดการพารามิเตอร์ที่แบ่งปันกันเป็นเวกเตอร์และเมทริกซ์ เฟรมเวิร์กนี้รองรับการสื่อสารข้อมูลแบบ asynchronous ระหว่าง node และสามารถปรับขนาดได้ตามความยืดหยุ่น (Elastic scalability) พร้อมการป้องกันความผิดพลาด (Fault tolerance) งานวิจัยแสดงผลการทดลองบนข้อมูลขนาดหลาย petabytes ที่ประกอบด้วยตัวอย่างและพารามิเตอร์หลายพันล้าน การจัดการ Distributed Optimization และ Inference จำเป็นต่อการแก้ปัญหาด้าน Machine Learning ขนาดใหญ่ เนื่องจากความซับซ้อนของโมเดลที่เพิ่มขึ้นตามข้อมูล งานนี้เน้นความท้าทายด้านการเข้าถึงพารามิเตอร์, การซิงโครไนซ์, ความล่าช้าในระบบ และการรองรับความผิดพลาดในสภาพแวดล้อมที่ใช้ loud โดยเฟรมเวิร์กนี้จะช่วยแก้ปัญหาดังกล่าวได้

2. การประมวลผลแบบขนานด้วย GPU และการเพิ่มประสิทธิภาพอัลกอริทึม (Parallel Computing with GPUs and Algorithm Optimization)

   การประมวลผลแบบขนานด้วย GPU (Graphics Processing Unit) เป็นเทคโนโลยีที่สำคัญในการเพิ่มความเร็วและประสิทธิภาพของการประมวลผล โดยเฉพาะในงานที่มีความซับซ้อน เช่น การจำลองทางวิทยาศาสตร์, การประมวลผลภาพ (Image Processing) และงานด้าน AI เป็นต้น การพัฒนาอัลกอริทึมที่เหมาะสมกับสถาปัตยกรรมของ GPU จะช่วยให้สามารถใช้พลังการประมวลผลได้อย่างเต็มที่ ทำให้การดำเนินงานที่ต้องใช้ทรัพยากรจำนวนมากสามารถเสร็จสิ้นได้ในเวลาที่สั้นลง การวิจัยเกี่ยวกับ CUDA และ OpenCL ซึ่งเป็นแพลตฟอร์มสำหรับการเขียนโปรแกรมประมวลผลแบบขนานบน GPU จะช่วยให้ผู้พัฒนาโปรแกรมสามารถพัฒนา โปรแกรมที่ใช้ GPU ในการเร่งการประมวลผลและทำให้การทำงานในระบบ HPC มีประสิทธิภาพสูงขึ้น

   ตัวอย่างงานวิจัยที่เกี่ยวข้อง มีดังนี้

   - "**Optimizing matrix multiplication for a short-vector SIMD architecture – CELL processor**" โดย Jakub Kurzak, Wesley Alvaro และ Jack Dongarra
     งานวิจัยนี้เน้นไปที่การคูณเมทริกซ์ (Matrix Multiplication) ซึ่งเป็นหนึ่งในปฏิบัติการเชิงตัวเลขที่สำคัญในวิชาพีชคณิตเชิงเส้น และใช้เป็นพื้นฐานของอัลกอริทึมหลายอย่าง เช่น การแก้ระบบสมการเชิงเส้นและการคำนวณค่า eigenvalue งานวิจัยนี้มุ่งเน้นการใช้งานประมวลผลของ CELL processor ซึ่งมีประสิทธิภาพสูงในด้านการประมวลผล floating point ความแม่นยำเดี่ยว (single precision floating point) โดย CELL processor ประกอบด้วย Synergistic Processing Elements (SPEs) ที่ใช้สถาปัตยกรรม SIMD (Single Instruction Multiple Data) สำหรับการดำเนินงานพร้อมกันหลายข้อมูล งานวิจัยนี้นำเสนอ kernel สำหรับการคูณเมทริกซ์บน SPE และบรรลุประสิทธิภาพถึง 99.80% ของขีดความสามารถสูงสุด (25.55 Gflop/s) งานนี้มีประโยชน์ต่อการประยุกต์ใช้ในการพัฒนาอัลกอริทึมที่เน้นประสิทธิภาพสูงในงานคำนวณเชิงตัวเลขและการคูณเมทริกซ์
   - "**GPU Computation in Bioinspired Algorithms: A Review**" โดย M.G. Arenas, Antonio Mora, Gustavo Romero และ Pedro A. Castillo

งานวิจัยนี้มีแนวคิดหลักคือการใช้วิธีการ Bioinspired Methods ซึ่งต้องการทรัพยากรในการประมวลผลสูง จึงได้เสนอการใช้ การประมวลผลแบบขนาน (Parallelization) โดยเฉพาะการใช้หน่วยประมวลผลกราฟิก (GPU) เพื่อช่วยลดเวลาในการประมวลผลและเพิ่มความแม่นยำของผลลัพธ์ เนื่องจาก GPU มีประสิทธิภาพสูงและมีต้นทุนต่ำจากการพัฒนาในอุตสาหกรรมเกมส์ การวิจัยนี้เน้นไปที่การพัฒนาอัลกอริทึมขนานโดยใช้ GPU ที่สามารถเข้าถึงได้ง่ายในคอมพิวเตอร์ทั่วไป และใช้แพลตฟอร์มซอฟแวร์ในการโปรแกรมคำสั่ง GPU เช่น CUDA และ OpenCL ในการนำไปประยุกต์ใช้ในด้านชีววิทยาคอมพิวเตอร์และชีวสารสนเทศ ผลการวิจัยนี้มีประโยชน์ต่อการประมวลผลทางวิทยาศาสตร์ที่ต้องการประสิทธิภาพสูงและต้นทุนต่ำในหลายสาขาวิชา

3. ความปลอดภัยและการรักษาความเป็นส่วนตัวในระบบ HPC (Security and Privacy in HPC Systems)

   ในระบบการประมวลผลสมรรถนะสูง (HPC) ความปลอดภัยและการรักษาความเป็นส่วนตัวถือเป็นปัจจัยสำคัญ เนื่องจากข้อมูลที่ผ่านการประมวลผลมักมีความละเอียดอ่อน เช่น ข้อมูลทางวิทยาศาสตร์หรือข้อมูลส่วนบุคคลที่ต้องได้รับการปกป้อง การพัฒนามาตรการความปลอดภัยใน HPC จึงเป็นสิ่งจำเป็น ไม่ว่าจะเป็นการเข้ารหัสข้อมูล (Encryption) เพื่อป้องกันการถูกดักฟังระหว่างการถ่ายโอน หรือการตรวจสอบสิทธิ์ (Authentication) เพื่อจำกัดการเข้าถึงให้เฉพาะผู้ที่ได้รับอนุญาต เทคนิคเหล่านี้ต้องถูกปรับให้เข้ากับสถาปัตยกรรมของ HPC เพื่อให้สามารถรักษาความปลอดภัยของข้อมูลได้อย่างมีประสิทธิภาพและครอบคลุม

   ตัวอย่างงานวิจัยที่เกี่ยวข้อง มีดังนี้

   - "**Enhancing High-Performance Computing (HPC) Security: A Comprehensive Review of Detection and Protection Strategies**" โดย S Koleini และ B Pahlevanzadeh

     งานวิจัยนี้มีแนวคิดหลักเกี่ยวกับการแก้ไขปัญหาด้านความปลอดภัยที่เพิ่มขึ้นในระบบคอมพิวเตอร์สมรรถนะสูง (HPC) เนื่องจากมีความต้องการในการใช้ HPC และการวิเคราะห์ข้อมูลมากขึ้นในหลากหลายสาขาวิทยาศาสตร์ งานวิจัยนี้ได้สำรวจและวิเคราะห์ปัญหาความปลอดภัยต่าง ๆ ที่พบในระบบ HPC โดยแบ่งกลยุทธ์ในการป้องกันออกเป็นสองประเภทหลัก: การตรวจจับ (Detection) และการป้องกัน (Protection) โดยใช้การวิเคราะห์ซอฟต์แวร์ทั้งแบบสถิต (Static Analysis) และแบบไดนามิก (Dynamic Analysis) นอกจากนี้ ยังมีการใช้ระบบเฝ้าระวังเพื่อตรวจจับและหยุดกิจกรรมที่เป็นอันตราย ซึ่งส่วนใหญ่เกี่ยวกับการโจมตีที่มุ่งทำลายความลับ (Confidentiality), ความสมบูรณ์ (Integrity), และความพร้อมใช้งาน (Availability) ของระบบ HPC งานวิจัยนี้นำเสนอแนวทางการป้องกัน เช่น การควบคุมการเข้าถึง (Access Control), การสุ่ม (Randomization), การรักษาความสมบูรณ์ของโฟลว์ควบคุม (Control Flow Integrity) และการทนทานต่อข้อผิดพลาด (Fault Tolerance) เพื่อช่วยเพิ่มประสิทธิภาพและความปลอดภัยของระบบ HPC

4. การเพิ่มประสิทธิภาพพลังงานในระบบ HPC (Energy Efficiency Optimization in HPC Systems)

การเพิ่มประสิทธิภาพพลังงานในระบบการประมวลผลสมรรถนะสูง (HPC) เป็นสิ่งสำคัญเพื่อให้สามารถรองรับการทำงานที่ซับซ้อนและประหยัดพลังงานได้อย่างยั่งยืน หนึ่งในวิธีที่สามารถนำมาใช้คือการวิจัยและพัฒนาอัลกอริทึมที่ช่วยลดการใช้พลังงาน เช่น Dynamic Voltage and Frequency Scaling (DVFS) ซึ่งช่วยปรับความถี่และแรงดันไฟฟ้าของโปรเซสเซอร์ให้เหมาะสมกับการทำงาน นอกจากนี้ การสำรวจการใช้พลังงานทดแทนและการพัฒนาระบบระบายความร้อนที่มีประสิทธิภาพยังเป็นอีกทางเลือกที่สำคัญ เพื่อช่วยลดผลกระทบด้านสิ่งแวดล้อมและลดค่าใช้จ่ายในการดำเนินงานของระบบ HPC

ตัวอย่างงานวิจัยที่เกี่ยวข้อง มีดังนี้

- "**Energy-Aware Scheduling for High-Performance Computing Systems: A Survey**" โดย Bartlomiej Kocot, Pawel Czarnul และ Jerzy Proficz

  งานวิจัยนี้มีแนวคิดเกี่ยวกับการทำ Scheduling Method ที่คำนึงถึงพลังงานในระบบคอมพิวเตอร์สมรรถนะสูง (HPC) ซึ่งเป็นการปรับปรุงที่สำคัญเนื่องจากปัญหาด้านค่าใช้จ่ายและสิ่งแวดล้อมในการใช้พลังงานของระบบ HPC โดยเน้นการเพิ่มประสิทธิภาพการประมวลผลผ่านการใช้พลังงานที่น้อยลง งานวิจัยนี้ได้รวบรวมวิธีการจัดตารางงาน (scheduling) ที่คำนึงถึงพลังงานในระบบ HPC หลากหลายประเภท รวมถึงการใช้เทคนิค Dynamic Voltage and Frequency Scaling (DVFS) และ Power Capping ซึ่งช่วยลดการใช้พลังงานและเพิ่มประสิทธิภาพในการคำนวณ นอกจากนี้ยังมีการวิเคราะห์อัลกอริทึมที่ใช้แก้ปัญหาการจัดตารางงาน (scheduling) ที่เกี่ยวข้องกับพลังงาน ทั้งใน CPU, GPU และคลัสเตอร์ที่มีลักษณะโฮโมจีเนียสและเฮเทอโรจีเนียส

- "**Energy-Aware High-Performance Computing: Survey of State-of-the-Art Tools, Techniques, and Environments**" โดย Pawel Czarnul, Jerzy Proficz และ Adam Krzywaniak

  งานวิจัยนี้มีแนวคิดเกี่ยวกับการจัดการพลังงานในระบบคอมพิวเตอร์สมรรถนะสูง (HPC) โดยเน้นการระบุและจำแนกวิธีการควบคุมพลังงานตามประเภทของระบบและอุปกรณ์ เช่น CPU, GPU, ระบบคลัสเตอร์ และระบบไฮบริด การเพิ่มประสิทธิภาพพลังงานครอบคลุมการจัดตารางงาน (scheduling), Dynamic Voltage and Frequency Scaling (DVFS), การจำกัดพลังงาน (power capping) และการใช้ API เช่น Intel RAPL และ NVIDIA NVML เพื่อควบคุมพลังงาน โดยเน้นการปรับปรุงซอฟต์แวร์และการใช้เครื่องมือสำหรับการจัดการพลังงาน งานวิจัยนี้ยังสรุปเครื่องมือและ API ที่ช่วยพยากรณ์และจำลองการใช้พลังงานในระบบ HPC และนำเสนอปัญหาที่ยังไม่ได้รับการแก้ไขเพื่อเป็นแนวทางสำหรับการวิจัยในอนาคต ผลลัพธ์จากงานวิจัยนี้ช่วยให้นักวิจัยและผู้ดูแลระบบสามารถพัฒนาวิธีการจัดการพลังงานที่มีประสิทธิภาพยิ่งขึ้นในระบบ HPC

5. การจัดการงานประมวลผลและการจัดลำดับงานในคลัสเตอร์ HPC (Job Management and Scheduling in HPC Clusters)

   การจัดการงานประมวลผลและการจัดลำดับงานในคลัสเตอร์ HPC เป็นกระบวนการที่สำคัญสำหรับการใช้ทรัพยากรอย่างเต็มประสิทธิภาพ การพัฒนาอัลกอริทึม Scheduling ที่มีความสามารถในการ

จัดลำดับงานอย่างมีประสิทธิภาพสามารถช่วยลดเวลารอคอยของงานและเพิ่มความสามารถในการประมวลผลได้อย่างมีประสิทธิภาพสูงสุด นอกจากนี้ การสร้างเครื่องมือสำหรับการจัดการคลัสเตอร์ (Cluster Management Tools) ที่สามารถปรับตัวตามภาระงานที่เปลี่ยนแปลงได้แบบไดนามิก จะช่วยให้ระบบสามารถปรับทรัพยากรให้สอดคล้องกับความต้องการของงานในแต่ละช่วงเวลา ลดการสูญเสียทรัพยากรที่ไม่ได้ใช้งาน และเพิ่มความยืดหยุ่นของระบบในการจัดการภาระงานที่ซับซ้อนและหลากหลาย

ตัวอย่างงานวิจัยที่เกี่ยวข้อง มีดังนี้

- "**SLURM: Simple Linux Utility for Resource Management**" โดย Morris A. Jette, Andy B. Yoo, และ Mark Grondona

  งานวิจัยนี้นำเสนอการพัฒนาและออกแบบระบบจัดการทรัพยากรสำหรับคลัสเตอร์คอมพิวเตอร์ที่เรียกว่า Simple Linux Utility Resource Management (SLURM) โดย SLURM ถูกพัฒนาขึ้นเพื่อรองรับการประมวลผลแบบขนานในคลัสเตอร์ขนาดใหญ่ที่มีหน่วยประมวลผลจำนวนมาก ระบบนี้ออกแบบให้มีความยืดหยุ่น สามารถปรับแต่งได้ตามความต้องการของคลัสเตอร์ที่มีขนาดและโครงสร้างที่แตกต่างกัน นอกจากนี้ยังเน้น ความทนทานต่อความผิดพลาด (fault-tolerance) และรองรับการขยายตัว (scalability) เพื่อให้สามารถใช้งานได้กับคลัสเตอร์ขนาดใหญ่ ระบบนี้ถูกสร้างขึ้นโดยใช้ภาษา C และสามารถปรับแต่งโมดูลต่าง ๆ ได้โดยใช้ปลั๊กอิน (plug-in mechanism) ซึ่งทำให้สามารถรองรับการเชื่อมต่อเครือข่ายที่หลากหลาย การวิจัยนี้จึงเป็นประโยชน์ต่อทั้งผู้ใช้และนักออกแบบระบบ เนื่องจากให้สภาพแวดล้อมการจัดการงานขนานที่เรียบง่ายแต่มีประสิทธิภาพสูง

6. การพัฒนาเครือข่ายความเร็วสูงและโปรโตคอลสำหรับ HPC (Development of High-Speed Networks and Protocols for HPC)

   การพัฒนาเครือข่ายความเร็วสูงและโปรโตคอลสำหรับ HPC เป็นองค์ประกอบสำคัญในการเพิ่มประสิทธิภาพการถ่ายโอนข้อมูลขนาดใหญ่ ด้วยการวิจัยและออกแบบโปรโตคอลเครือข่ายที่มีประสิทธิภาพสูง เช่น InfiniBand และ Remote Direct Memory Access (RDMA) ทำให้สามารถลดความหน่วงเวลาในการถ่ายโอนข้อมูลและเพิ่มประสิทธิภาพในการประมวลผล นอกจากนี้ การพัฒนาเทคนิคการบีบอัดข้อมูลและการถ่ายโอนข้อมูลที่สอดคล้องกับสถาปัตยกรรม HPC ยังเป็นวิธีที่ช่วยลดการใช้แบนด์วิธและเพิ่มความรวดเร็วในการถ่ายโอนข้อมูล ซึ่งจะเป็นการสนับสนุนงานวิจัยที่ใช้การประมวลผลแบบขนานและการวิเคราะห์ข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพมากขึ้น

7. การประยุกต์ใช้ HPC ในการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analytics)

   การประยุกต์ใช้ HPC ในการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analytics) มีความสำคัญในการประมวลผลข้อมูลจำนวนมหาศาลที่มีความซับซ้อน การรวม HPC กับเทคโนโลยี Big Data จะช่วยเพิ่มความสามารถในการจัดการกับข้อมูลจำนวนมากผ่านการประมวลผลแบบขนานและการกระจายงานให้กับเครื่องหลายตัวพร้อมกัน ข้อเสนอแนะคือควรมีการวิจัยเพิ่มเติมเพื่อรวม HPC เข้ากับแพลตฟอร์ม

Big Data เพื่อรองรับงานวิเคราะห์ข้อมูลที่ต้องใช้การคำนวณหนัก และพัฒนาแพลตฟอร์มที่ออกแบบมาเพื่อรองรับการประมวลผลที่มีประสิทธิภาพมากขึ้นสำหรับการวิเคราะห์ข้อมูลขนาดใหญ่

6 ข้อสรุปและข้อเสนอแนะ (Conclusion and Recommendations)

การพัฒนา High-Performance Computing (HPC) ในประเทศไทยมีความสำคัญอย่างยิ่งในการเสริมสร้างศักยภาพทางวิทยาศาสตร์และเทคโนโลยี โดยบทความนี้ได้นำเสนอการพัฒนาระบบ HPC ในช่วง 25 ปีที่ผ่านมา รวมถึงสรุปข้อมูลที่มีอยู่ในปัจจุบัน เช่น การใช้ประโยชน์ (Application) จากแกนประมวลผลกว่า 54,838 cores และพื้นที่จัดเก็บข้อมูล 21 PB ในระบบประเทศไทย ในปี 2022 ระบบ HPC เหล่านี้ถูกนำไปใช้ในหลากหลายงานวิจัย เช่น การประยุกต์ใช้ใน AI, การจำลองทางวิทยาศาสตร์, และการวิเคราะห์ข้อมูลขนาดใหญ่ เป็นต้น อย่างไรก็ตามยังมีความท้าทายในการพัฒนาโครงสร้างพื้นฐานของ HPC เช่น ความต้องการด้านการรับส่งข้อมูลขนาดใหญ่และการบริหารจัดการทรัพยากรเพื่อให้รองรับปริมาณงานที่เพิ่มขึ้นในอนาคต

เพื่อนำ HPC มาใช้ให้เกิดประโยชน์สูงสุดในประเทศไทย ทางผู้เขียนมีข้อเสนอแนะสำหรับการวิจัยในอนาคต เพื่อส่งเสริม HPC ในประเทศไทยควรมุ่งเน้นไปที่การพัฒนาอัลกอริทึมและระบบที่สามารถใช้ประโยชน์จากสถาปัตยกรรม HPC ได้อย่างมีประสิทธิภาพ รวมถึงการวิจัยด้านการประยุกต์ใช้ HPC ในการฝึกสอนโมเดล AI ขนาดใหญ่ การเพิ่มประสิทธิภาพการประมวลผลแบบขนานด้วย GPU และการจัดการพลังงานในระบบ HPC เพื่อลดการใช้พลังงาน นอกจากนี้ ควรพัฒนาเครือข่ายความเร็วสูงและโปรโตคอลที่สามารถรองรับการถ่ายโอนข้อมูลขนาดใหญ่ได้อย่างรวดเร็ว เพื่อเตรียมพร้อมสำหรับเทคโนโลยีในอนาคต

เอกสารอ้างอิง

1. V. Varavithya and S. Prueksaaroon, “A Survey of High Performance Computing (HPC) Infrastructure in Thailand,” in *ECTI-CIT Transactions*, Bangkok, Thailand, 2023, pp. 255–264.

2. “Minicomputer,” *Wikipedia*, Available: https://en.wikipedia.org/wiki/Minicomputer. [Accessed: Sep. 17, 2024].

3. “Mainframe computer,” *Wikipedia*, Available: https://en.wikipedia.org/wiki/Mainframe_computer. [Accessed: Sep. 17, 2024].

4. “Vector processor,” *Wikipedia*, Available: https://en.wikipedia.org/wiki/Vector_processor. [Accessed: Sep. 17, 2024].

5. “Supercomputer,” *Wikipedia*, Available: https://en.wikipedia.org/wiki/Supercomputer. [Accessed: Sep. 17, 2024].

6. “Massively parallel,” *Wikipedia*, Available: https://en.wikipedia.org/wiki/Massively_parallel. [Accessed: Sep. 17, 2024].

7. “IBM Blue Gene,” *Wikipedia*, Available: https://en.wikipedia.org/wiki/IBM_Blue_Gene. [Accessed: Sep. 17, 2024].

8. “Tianhe-2,” *Wikipedia*, Available: https://en.wikipedia.org/wiki/Tianhe-2. [Accessed: Sep. 17, 2024].

9. “Computer cluster,” *Wikipedia*, Available: https://en.wikipedia.org/wiki/Computer_cluster. [Accessed: Sep. 17, 2024].

10. “APEX research project,” *CMKL University*, Available: https://www.cmkl.ac.th/research/apex. [Accessed: Sep. 17, 2024].

11. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2012. Available: https://papers.nips.cc/paper_files/paper/2012/hash/6aca97005c68f1206823815f66102863-Abstract.html. [Accessed: Sep. 17, 2024].

12. T. Chen, B. Xu, C. Zhang, and C. Guestrin, “Training deep nets with sublinear memory cost,” in *Proceedings of the 27th Conference on Neural Information Processing Systems*, vol. 2, 2014, pp. 1548–1556. Available: https://dl.acm.org/doi/10.5555/2685048.2685095. [Accessed: Sep. 17, 2024].

13. B. He and N. K. Govindaraju, “Efficient algorithms for molecular dynamics simulations on graphics processing units,” *Journal of Computational Science*, vol. 4, no. 3, pp. 345–352, 2009. Available: https://www.sciencedirect.com/science/article/abs/pii/S016781910900012X. [Accessed: Sep. 17, 2024].

14. R. Ruiz, A. F. Rossi, and C. M. Cuadros, “GPU computation in bioinspired algorithms: A review,” *ResearchGate*, 2011. Available: https://www.researchgate.net/publication/225210354_GPU_Computation_in_Bioinspired_Algorithms_A_Review. [Accessed: Sep. 17, 2024].

15. A. Z. Kouzaei and M. Taghipour, “A new parallel algorithm for matrix multiplication using dynamic multithreading,” *Journal of Computing and Security*, vol. 7, no. 3, pp. 45–52, 2020. Available: https://www.jdcs.ir/article_201350.html. [Accessed: Sep. 17, 2024].

16. K. Iskra, P. Balaji, Y. Sabharwal, and R. Thakur, “Energy-aware high-performance computing: Survey of state-of-the-art tools, techniques, and environments,” *ResearchGate*, 2019. Available: https://www.researchgate.net/publication/332647237_Energy-Aware_High-Performance_Computing_Survey_of_State-of-the-Art_Tools_Techniques_and_Environments. [Accessed: Sep. 17, 2024].

17. A. Chudnovskyi, P. Sorokin, and O. Zholtkevych, “High-performance computing for scalable data analysis and processing,” *Proceedings of the Ukrainian Conference on Computer Science*, vol. 2, pp. 12–17, 2019. Available: https://ouci.dntb.gov.ua/en/works/9j6qjMg9/. [Accessed: Sep. 17, 2024].

18. B. Schmidt, “Efficient algorithms for parallel computing on multi-core processors,” in *Proceedings of the Lecture Notes in Computer Science*, 2005, pp. 45–60. Available: https://link.springer.com/chapter/10.1007/10968987_3. [Accessed: Sep. 17, 2024].