การตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา

นายอภิชิต หาญบรรจง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

SARIMA BASED NETWORK BANDWIDTH ANOMALY DETECTION

Mr.Aphichit Hanbanchong

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2011

หัวข้อวิทยานิพนธ์	การตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา
โดย	นายอภิชิต หาญบรรจง
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.เกริก ภิรมย์โสภา

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

…….. คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

…… ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐวุฒิ หนูไพโรจน์)

…….. อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.เกริก ภิรมย์โสภา)

…… กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

…… กรรมการภายนอกมหาวิทยาลัย
(ดร.พงศ์ธวัช ชีพพิมลชัย)

อภิชิต หาญบรรจง : การตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา. (SARIMA BASED NETWORK BANDWIDTH ANOMALY DETECTION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร.เกริก ภิรมย์โสภา, 28หน้า.

แบนด์วิดท์ของเครือข่ายถือว่าเป็นทรัพยากรที่สำคัญของระบบคอมพิวเตอร์ ปัจจุบันจึงมีการใช้ระบบตรวจหาการบุกรุก (Intrusion detection system) เพื่อตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่าย อย่างไรก็ตามระบบตรวจหาการบุกรุกที่มีประสิทธิภาพต้องสามารถตรวจหาความผิดปกติของเครือข่ายได้โดยเกิดผลบวกลวง (False positive) น้อย วิธีหนึ่งที่สามารถนำมาใช้ในการพยากรณ์อัตราการใช้งานแบนด์วิดท์ของเครือข่ายได้ดีคือการใช้อนุกรมเวลาควบคู่กับค่าขีดสุด (Threshold) การวิจัยนี้นำเสนอการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา (SARIMA) โดยใช้ค่าขีดสุดเท่ากับ 8.5 เปอร์เซนต์ของค่าสูงสุดของอัตราการใช้งานแบนด์วิดท์ของเครือข่ายของแต่ละวัน ซึ่งผลลัพธ์ของการใช้วิธีนี้ทำให้สามารถตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายได้อย่างมีประสิทธิภาพและเกิดผลบวกลวงเพียง 3.57% เมื่อเทียบกับการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้อาริมา (ARIMA) จะได้ว่าการใช้ซาริมาจะเกิดผลบวกลวงน้อยกว่า และใช้ค่าขีดสุดที่ต่ำกว่า

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....
สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....
ปีการศึกษา 2554

5370522421 : MAJOR COMPUTER ENGINEERING

KEYWORDS: FORECASTING / INTRUSION DETECTION / SARIMA / SECURITY

APHICHIT HANBANCHONG : SARIMA BASED NETWORK BANDWIDTH ANOMALY DETECTION. ADVISOR : ASST.PROF.KRERK PIROMSOPA, Ph.D., 28 pp.

Network bandwidth is considered a valuable resource in most computer systems. To precisely detect network anomalies (with a few false alarms), an intrusion detection system requires reliable methods. A potential solution in predicting network bandwidth usage is using a time-series model with a threshold. This paper proposes a network anomaly detection technique based on SARIMA, a time-series model, to capture seasonal behavior of bandwidth usage of most networks. Our proposed SARIMA based anomaly detection is capable of detecting network bandwidth anomalies effectively when a threshold equals to 8.5 percent of maximum bandwidth in a day. Our result yields 3.57 percent of false alarms. We concluded that SARIMA is a better instrumental tool for intrusion detection comparing to ARIMA.

Department : Computer Engineering Student's Signature : .....

Field of Study : Computer Engineering Advisor's Signature : .....

Academic Year : 2011

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์อย่างยิ่งของผู้ช่วยศาสตราจารย์ ดร.เกริก ภิรมย์โสภา อาจารย์ที่ปรึกษา ซึ่งท่านได้ให้คำแนะนำตลอดการวิจัย และให้ความช่วยเหลือในการแก้ปัญหาต่าง ๆ เป็นอย่างดี จนทำให้การวิจัยในครั้งนี้สำเร็จลุล่วงด้วยดี

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.ณัฐวุฒิ หนูไพโรจน์ ผู้ช่วยศาสตราจารย์ ดร.โชติ รัตน์ รัตนามหัทธนะ และดร.พงศ์ธวัช ชีพพิมลชัย กรรมการสอบวิทยานิพนธ์ ที่กรุณาเสียสละเวลาให้คำแนะนำ ตรวจสอบ และแก้ไขวิทยานิพนธ์ฉบับนี้

ท้ายที่สุด ผู้เสนอวิทยานิพนธ์ขอขอบคุณเพื่อนๆ ครอบครัว และท่านอื่น ๆ ที่มิได้กล่าวชื่อไว้ ณ ที่นี้ ที่คอยให้กำลังใจและสนับสนุนให้วิทยานิพนธ์นี้สำเร็จได้ด้วยดี

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

ในบทนำนี้จะแบ่งเป็น 8 หัวข้อย่อย กล่าวถึงความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย ขอบเขตของการวิจัย คำจำกัดความที่ใช้ในการวิจัย ประโยชน์ที่คาดว่าจะได้รับ วิธีดำเนินการวิจัย ลำดับขั้นตอนในการเสนอผลการวิจัย และผลงานที่ตีพิมพ์จากวิทยานิพนธ์ ตามลำดับ ดังนี้

1.1 ความเป็นมาและความสำคัญของปัญหา

ความปลอดภัยและความเชื่อถือได้ของระบบเครือข่ายเป็นเรื่องที่สำคัญ ถึงแม้ว่าในปัจจุบันมีปัจจัยหลากหลายประเภทที่ก่อให้เกิดความเสียหายกับระบบเครือข่ายได้ แต่ปัจจัยหนึ่งที่สำคัญมากคือการโจมตี และการโจมตีที่ป้องกันได้ยากคือการโจมตีเพื่อให้เกิดการหยุดการให้บริการ (Denial-of-service attack) การโจมตีประเภทนี้คือ การพยายามทำให้ทรัพยากรของคอมพิวเตอร์หรือระบบเครือข่ายไม่สามารถให้บริการแก่ผู้ใช้ที่ต้องการได้ [1] โดยเฉพาะทรัพยากรแบนด์วิดท์ ผู้โจมตีจะส่งแพ็คเก็ตจำนวนมากมาที่ระบบเครือข่ายเป้าหมายเพื่อให้เกินความสามารถของระบบเครือข่ายเป้าหมายที่จะรองรับได้ ตัวอย่างของการโจมตีประเภทนี้ได้แก่ การโจมตีในวันที่ 6 สิงหาคม 2552 ซึ่งมีเป้าหมายคือทวิตเตอร์ ส่งผลให้ทวิตเตอร์ไม่สามารถให้บริการได้เป็นเวลาหลายชั่วโมง และเกิดผลกระทบให้ทวิตเตอร์ใช้เวลาในการโหลดนานขึ้น [2]

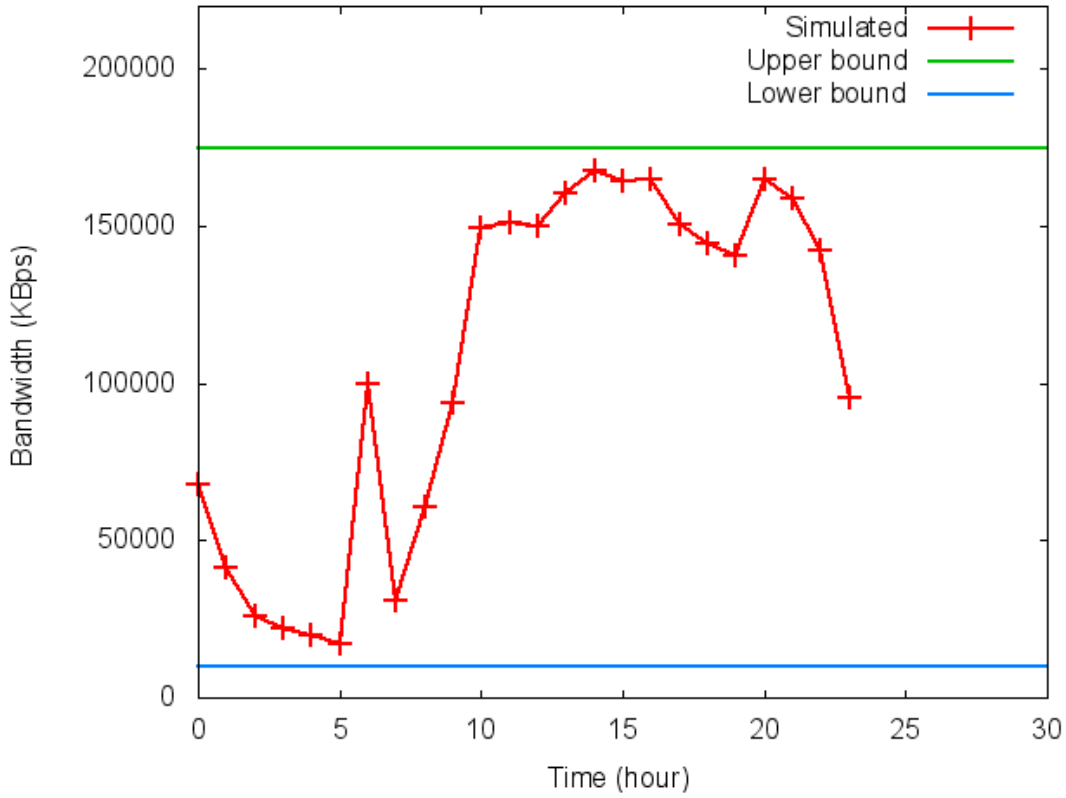
ระบบตรวจหาการบุกรุก (Intrusion detection system) เป็นระบบหนึ่งที่สามารถนำมาใช้เพื่อแก้ปัญหาเหล่านี้ได้ โดยระบบตรวจหาการบุกรุกจะตรวจหาการโจมตีและแจ้งเตือนไปยังผู้ดูแลระบบเพื่อให้ตรวจสอบพฤติกรรมน่าสงสัย ข้อดีของการใช้ระบบตรวจหาการบุกรุกคือ ระบบนี้สามารถตรวจหาการโจมตีให้มากที่สุดเท่าที่จะมากได้ แต่อย่างไรก็ตามระบบตรวจหาการบุกรุกมีข้อเสียคือ ไม่สามารถป้องกันการโจมตีได้ด้วยตัวเองต้องให้ผู้ดูแลระบบสร้างกฎขึ้นมาต่างหาก

การตรวจหาการบุกรุกแบ่งออกได้เป็น 2 ประเภทได้แก่ การตรวจหาการบุกรุกโดยใช้ลายมือชื่อ (Signature-based detection) และการตรวจหาการบุกรุกโดยใช้ความผิดปกติ (Anomaly-based detection) [3] การตรวจหาการบุกรุกโดยใช้ลายมือชื่อนั้นจะต้องมีการศึกษารูปแบบการโจมตี แล้วกำหนดลายมือชื่อขึ้นมา แต่การตรวจหาการบุกรุกโดยใช้ความผิดปกติจะใช้การวิเคราะห์กิจกรรมภายในระบบ แล้วจัดกลุ่มกิจกรรมเหล่านั้นว่าเป็นกิจกรรมปกติหรือกิจกรรมผิดปกติโดยใช้วิทยาการศึกษาสำนึก (Heuristics) หรือกฎต่างๆ ซึ่งการตรวจหาการบุกรุกประเภท

การโจมตีโดยใช้แบนด์วิดท์ของเครือข่ายจำเป็นต้องนำอัตราการใช้แบนด์วิดท์ของเครือข่ายมาวิเคราะห์เพื่อให้สามารถตรวจหาการโจมตีได้อย่างแม่นยำ

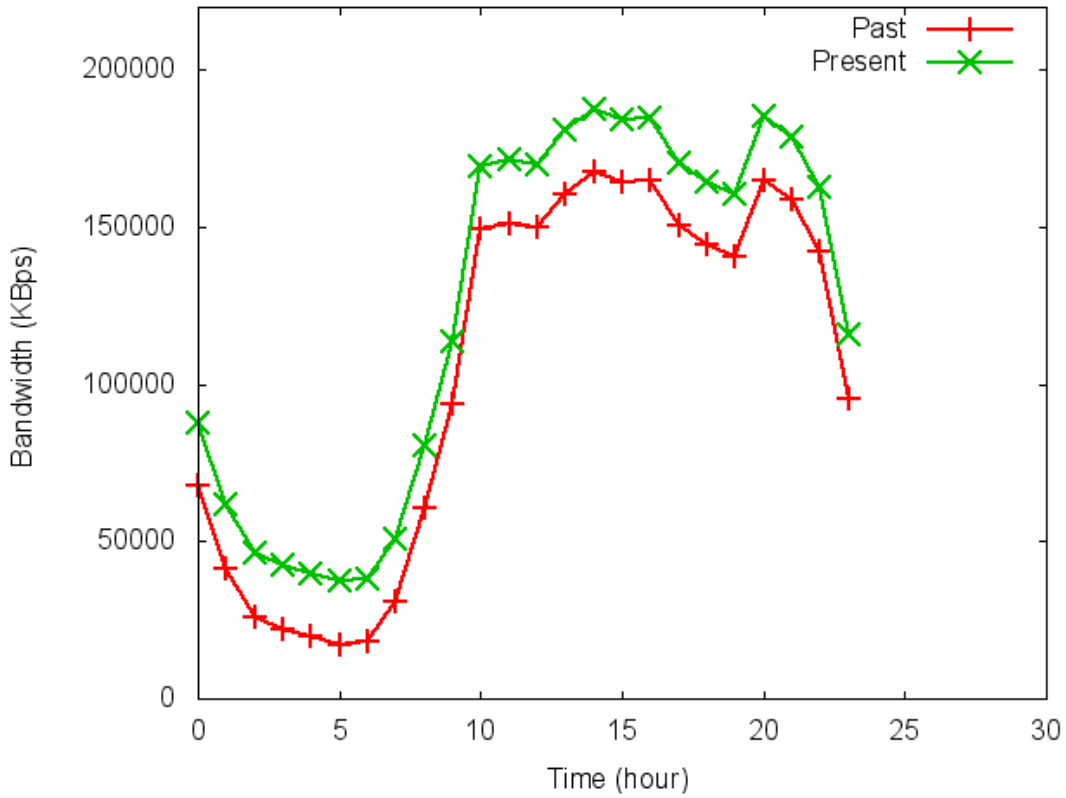
การตรวจหาความผิดปกติแบนด์วิดท์โดยทั่วไปมี 2 วิธี ได้แก่

1. การตรวจหาความผิดปกติแบนด์วิดท์โดยใช้ขอบเขตบนและขอบเขตล่าง วิธีนี้มีปัญหาคือไม่สามารถตรวจหาความผิดปกติแบนด์วิดท์ที่มีอัตราการใช้แบนด์วิดท์อยู่ภายในขอบเขตบนและขอบเขตล่างได้ ดังภาพที่ 1 ซึ่งมีความผิดปกติเกิดขึ้น ณ ชั่วโมงที่ 6



ภาพที่ 1 การตรวจหาความผิดปกติโดยใช้ขอบเขตบนและขอบเขตล่าง

2. การตรวจหาความผิดปกติแบนด์วิดท์โดยใช้สถิติ เป็นการนำข้อมูลอัตราการใช้แบนด์วิดท์ในอดีตมาเปรียบเทียบกับปัจจุบันเพื่อหาค่าความผิดพลาด ถ้าค่าความผิดพลาดเกินค่าขีดสุดถือว่ามีความผิดปกติเกิดขึ้น วิธีนี้มีปัญหาคือระบบเครือข่ายมีการเปลี่ยนแปลงอยู่ตลอดเวลา จำนวนผู้ใช้งานในแต่ละปีอาจจะไม่เท่ากัน ทำให้มีค่าความผิดพลาดสูง ดังภาพที่ 2



ภาพที่ 2 การตรวจหาความผิดปกติโดยใช้สถิติ

แบบจำลองอนุกรมเวลาอาริมาจึงถูกนำมาใช้ในการตรวจหาความผิดปกติของเครือข่ายในงานวิจัย [4] และได้ผลลัพธ์ที่ดี แต่อย่างไรก็ตามอัตราการใช้แบนด์วิดท์ของเครือข่ายโดยทั่วไปแล้วจะเหมาะสมต่อการพยากรณ์ด้วยแบบจำลองอนุกรมเวลาซาริมา [5] ดังนั้นการวิจัยนี้จึงนำเสนอวิธีการตรวจหาความผิดปกติของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาซาริมา

1.2 วัตถุประสงค์ของการวิจัย

การวิจัยนี้มีวัตถุประสงค์ ดังนี้

1. เพื่อออกแบบและพัฒนาวิธีการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาซาริมา
2. เพื่อทดสอบและเปรียบเทียบประสิทธิภาพระหว่างการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาซาริมาและแบบจำลองอนุกรมเวลาอาริมา

1.3 ขอบเขตของการวิจัย

ขอบเขตของการวิจัยถูกกำหนดไว้ ดังนี้

1. ออกแบบและทดสอบระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาซาริมาเปรียบเทียบกับแบบจำลองอนุกรมเวลาอาริมา โดยสนใจเฉพาะหน่วยวิเคราะห์อัตราการใช้แบนด์วิดท์เท่านั้น
2. ตรวจหาความผิดปกติในเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาซาริมานี้ตรวจหาเฉพาะความผิดปกติของอัตราการใช้แบนด์วิดท์เท่านั้น

1.4 คำจำกัดความที่ใช้ในการวิจัย

คำจำกัดความที่ใช้ในการวิจัย มีดังต่อไปนี้

- อัตราการใช้แบนด์วิดท์ หมายถึง ปริมาณการใช้แบนด์วิดท์ต่อหนึ่งชั่วโมง

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย ได้แก่

1. เข้าใจปัญหาและวิธีการแก้ไขปัญหาความผิดปกติแบนด์วิดท์ของเครือข่ายประเภทการใช้ทรัพยากรแบนด์วิดท์
2. ได้วิธีการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายประเภทการใช้ทรัพยากรแบนด์วิดท์ที่มีความแม่นยำสูงและมีผลบวกลวง (False positive) เกิดขึ้นน้อย
3. สามารถนำความรู้จากงานวิจัยนี้ไปประยุกต์ใช้ต่อไปได้

1.6 วิธีดำเนินการวิจัย

วิธีดำเนินการวิจัย ถูกแบ่งเป็น 5 ขั้นตอน ดังนี้

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง
2. ออกแบบระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่าย
3. สร้างระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่าย
4. ทดสอบประสิทธิภาพระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่าย
5. สรุปผลการวิจัยและจัดทำวิทยานิพนธ์

1.7 ลำดับขั้นตอนในการเสนอผลการวิจัย

วิทยานิพนธ์นี้แบ่งเนื้อหาออกเป็น 5 บท ดังต่อไปนี้ บทที่ 1 เป็นบทนำ ซึ่งกล่าวถึงความเป็นมาและความสำคัญของปัญหา รวมถึงวัตถุประสงค์ของการวิจัย บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง บทที่ 3 กล่าวถึงหลักการของระบบตรวจหาความผิดปกติแบนด์วิดท์ของ

เครือข่ายโดยใช้ซาริมา บทที่ 4 กล่าวถึงการทดสอบประสิทธิภาพของระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา และบทที่ 5 กล่าวถึงสรุปผลการวิจัย

1.8 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตอบรับให้ตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง “SARIMA Based Network Bandwidth Anomaly Detection” [6] โดยนายอภิชิต หาญบรรจง และผู้ช่วยศาสตราจารย์ ดร. เกริก ภิรมย์โสภา, ในงานประชุมวิชาการ “The Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE’12)” ณ มหาวิทยาลัยหอการค้าไทย จังหวัดกรุงเทพฯ ระหว่างวันที่ 30 พฤษภาคม – 1 มิถุนายน 2555

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงแนวคิดและทฤษฎี รวมทั้งเอกสารและงานวิจัยที่เกี่ยวข้อง ดังนี้

2.1 แนวคิดและทฤษฎีที่เกี่ยวข้อง

แนวคิดและทฤษฎีที่เกี่ยวข้องมีดังนี้

2.1.1 แบบจำลองอนุกรมเวลาอาริมา (ARIMA)

แบบจำลองอนุกรมเวลาอาริมาถูกพัฒนาขึ้นโดย Box และ Jenkins [7] แบบจำลองนี้ได้รับความนิยมสูงในหลายสาขาเนื่องจากมีความสามารถและความยืดหยุ่นสูง [8] แบบจำลองอนุกรมเวลาอาริมาสามารถถูกเขียนได้ในรูป $\mathrm{ARIMA(p,d,q)}$ ซึ่ง p d และ q หมายถึงอันดับของ Autoregressive Differencing และ Moving average ตามลำดับ [9] กำหนดให้ $\{\mathrm{X_t}: \mathrm{t} = \cdots, -1,0,1, \cdots\}$ เป็นโปรเซสของ $\mathrm{ARIMA(p,d,q)}$ จะได้ว่าสมการของ $\mathrm{ARIMA(p,d,q)}$ คือ

$$\phi_{\mathrm{p}}(\mathrm{B})\nabla^{\mathrm{d}}\mathrm{X_t} = \theta_{\mathrm{q}}(\mathrm{B})\mathrm{a_t} \tag{1}$$

จากสมการที่ 1 ตัวแปรต่างๆของสมการมีดังนี้

- B เป็นตัวดำเนินการ Backward-shift ซึ่ง $\mathrm{BX_t} = \mathrm{X_{t-1}}$
- $\nabla = 1 - \mathrm{B}$ เป็นตัวดำเนินการ Differencing
- $\{\mathrm{a_t}: \mathrm{t} = \cdots, -1,0,1, \ldots\}$ เป็นสัญญาณรบกวนขาว (White Noise) $\mathrm{WN}(0, \sigma^2)$ ซึ่งมีค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น σ^2
- $\phi_{\mathrm{p}}(\mathrm{B}) = 1 - \phi_1\mathrm{B} - \phi_2\mathrm{B}^2 - \cdots - \phi_{\mathrm{p}}\mathrm{B}^{\mathrm{p}}$
- $\theta_{\mathrm{q}}(\mathrm{B}) = 1 - \theta_1\mathrm{B} - \theta_2\mathrm{B}^2 - \cdots - \theta_{\mathrm{q}}\mathrm{B}^{\mathrm{q}}$

2.1.2 แบบจำลองอนุกรมเวลาซาริมา (SARIMA)

แบบจำลองอนุกรมเวลาซาริมาเป็นแบบจำลองที่เพิ่มส่วนฤดูกาล (Seasonal part) ให้กับแบบจำลองอนุกรมเวลาอาริมา และสามารถถูกเขียนได้ในรูป $\mathrm{ARIMA(p,d,q)} \times \mathrm{(P,D,Q)_s}$ ซึ่ง P D Q และ s หมายถึงอันดับของ Seasonal autoregressive Seasonal differencing Seasonal moving average และ คาบฤดูกาล (Seasonal period) ตามลำดับ [9] สมการของ $\mathrm{ARIMA(p,d,q)} \times \mathrm{(P,D,Q)_s}$ คือ

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D X_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{2}$$

จากสมการที่ 2 ตัวแปรต่างๆของสมการมีดังนี้

- B เป็นตัวดำเนินการ Backward-shift ซึ่ง $BX_t = X_{t-1}$ และ $B^sX_t = X_{t-s}$
- $\nabla = 1 - B$ เป็นตัวดำเนินการ Differencing ซึ่ง $\nabla_s = 1 - B^s$
- $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$
- $\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$
- $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{s2} - \cdots - \Phi_P B^{sP}$
- $\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{s2} - \cdots - \Theta_Q B^{sQ}$

2.1.3 Akaike information criterion (AIC)

AIC คือการวัดความเหมาะสมของแบบจำลองทางสถิติที่ถูกประมาณ โดยใช้แนวคิดของเอนโทรปี เพื่อวัดการสูญหายของข้อมูลเมื่อใช้แบบจำลองนั้น โดยแบบจำลองที่มีค่า AIC น้อยที่สุดคือแบบจำลองที่ดีที่สุด [10]

2.1.4 Bayesian information criterion (BIC)

BIC คือบรรทัดฐานในการเลือกแบบจำลองภายในกลุ่มของแบบจำลองที่มีการปรับค่าตัวแปรต่างๆ โดย BIC มีความเกี่ยวข้องกับ AIC มาก แต่ BIC จะมีการให้โทษสำหรับตัวแปรที่เพิ่มขึ้นมากกว่า

2.1.5 Maximum likelihood estimation (MLE)

MLE คือวิธีทางสถิติที่นิยมใช้ในการปรับแบบจำลองทางสถิติให้เหมาะสมกับข้อมูลและหาค่าประมาณตัวแปรของแบบจำลองนั้น โดย MLE จะเลือกค่าตัวแปรที่ทำให้ฟังก์ชันภาวะน่าจะเป็น (Likelihood function) มีค่ามากที่สุด [11]

2.1.6 Minimum mean-square error (MMSE)

MMSE คือการประมาณทางสถิติโดยทำให้ Mean-square error (MSE) ซึ่งเป็นค่าที่นิยมนำมาใช้วัดคุณภาพของการประมาณมีค่าน้อยที่สุด

2.1.7 Root-mean-square error (RMSE)

RMSE คือ วิธีการในการวัดความแตกต่างระหว่างค่าจริงกับค่าที่ได้จากพยากรณ์ที่มีความแม่นยำสูง โดยมีสมการดังนี้

$$\mathrm{RMSE}(\hat{\mathrm{x}}) = \sqrt{\mathrm{E}((\hat{\mathrm{x}} - \mathrm{x})^2)} \tag{3}$$

จากสมการที่ 3 ตัวแปรต่างๆของสมการมีดังนี้

- x เป็นค่าจริง
- $\hat{\mathrm{x}}$ เป็นค่าที่ได้จากการทำนาย
- $\mathrm{E}(\mathrm{x})$ เป็นค่าคาดหวังของ x

2.1.8 Normalized root-mean-square error (NRMSE)

NRMSE คือ RMSE ที่ถูกทำให้อยู่ในรูปปกติ (Normalize) โดยมีสมการดังนี้

$$\mathrm{NRMSE}(\hat{\mathrm{x}}) = \frac{\mathrm{RMSE}(\hat{\mathrm{x}})}{\mathrm{x_{max}} - \mathrm{x_{min}}} \tag{4}$$

จากสมการที่ 4 ตัวแปรต่างๆของสมการมีดังนี้

- $\mathrm{x_{max}}$ เป็นค่า x ที่มากที่สุด
- $\mathrm{x_{min}}$ เป็นค่า x ที่น้อยที่สุด

2.2 **งานวิจัยที่เกี่ยวข้อง**

งานวิจัยที่เกี่ยวข้องมีดังนี้

2.2.1 ARIMA Based Network Anomaly Detection [4]

งานวิจัยของ Yaacob และ Tan นำเสนอการตรวจหาความผิดปกติของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาอาริมา โดยพยายามพยากรณ์อัตราการใช้แบนด์วิดท์ที่จะเกิดขึ้นใน

ปัจจุบันจากอัตราการใช้แบนด์วิดท์ที่เกิดขึ้นในอดีต แล้วเปรียบเทียบความแตกต่างระหว่างอัตราการใช้แบนด์วิดท์ที่ทำนายไว้กับอัตราการใช้แบนด์วิดท์ที่เกิดขึ้นจริงว่ามีความแตกต่างกันเกินค่าขีดสุดหรือไม่ ถ้าเกินค่าขีดสุดจะถือว่ามีความผิดปกติ งานวิจัยนี้สามารถตรวจหาการโจมตีเพื่อให้เกิดการหยุดให้บริการได้อย่างมีประสิทธิภาพ แต่มีข้อจำกัดอยู่สองด้านคือ ไม่สามารถใช้งานได้กับเครือข่ายที่มีอัตราการใช้แบนด์วิดท์น้อยกว่าหนึ่งเมกะไบต์ต่อวินาที และมีผลบวกลวงเกิดขึ้น

2.2.2 Time Series Model for Internet Traffic [5]

งานวิจัยของ Basu และ Mukherjee พยายามสร้างแบบจำลองอนุกรมเวลาที่เหมาะสมสำหรับการจราจรของเครือข่ายอินเตอร์เน็ต และได้ข้อสรุปว่าการจราจรของเครือข่ายอินเตอร์เน็ตจำนวนมากสามารถพยากรณ์ได้ด้วยแบบจำลองอนุกรมเวลาซาริมา โดยกรณีส่วนใหญ่จะมีคาบฤดูกาลที่ชัดเจน

2.2.3 Wireless Traffic Modeling and Prediction Using Seasonal ARIMA Models [9]

Shu, Yu และ Liu นำงานวิจัย [5] มาใช้ในการนำเสนอการสร้างแบบจำลองและพยากรณ์การจราจรของเครือข่ายไร้สายโดยใช้แบบจำลองอนุกรมเวลาซาริมา และนำเสนออัลกอริทึมการสร้างแบบจำลองอนุกรมเวลาซาริมา ไว้ดังนี้

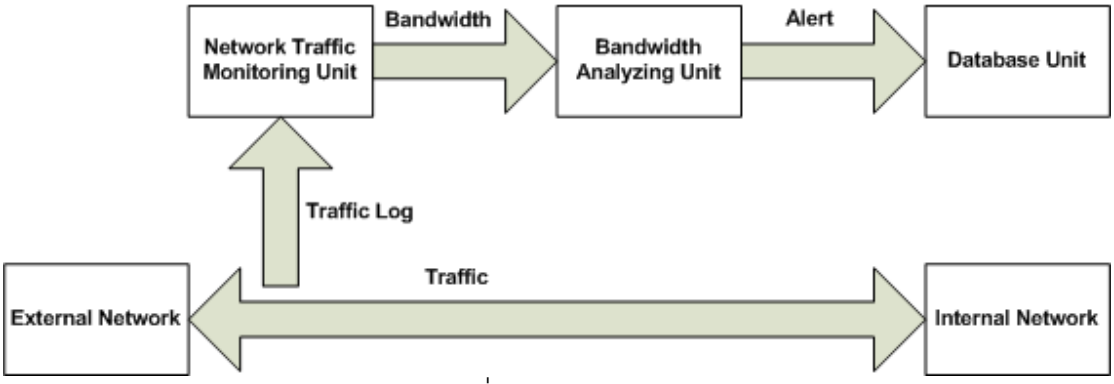
1. หาค่าของคาบฤดูกาล เช่น s_1 และ s_2 จากการวิเคราะห์สเปกตรัม
2. หาค่าประมาณของ d , D_1 และ D_2 จากการวิเคราะห์การจราจร แล้วทดสอบด้วยการทดสอบ ADF
3. ทำ Differencing ที่ X_t ตามสมการ $W_t = \nabla^d \nabla_{s_1}^{D_1} \nabla_{s_2}^{D_2} X_t$ เพื่อหาอนุกรมเวลาที่มีคุณสมบัติสเตชันนารี (Stationary series)
4. หาค่าอันดับ p q P_1 Q_1 P_2 และ Q_2 โดยการทดสอบแทนค่า 0 1 และ 2 ทุกรูปแบบ ซึ่ง p P_1 และ P_2 ไม่ควรเป็น 0 พร้อมกัน และ q Q_1 และ Q_2 ไม่ควรเป็น 0 พร้อมกัน จากนั้นเลือกรูปแบบการแทนค่าที่เหมาะสมที่สุดมาใช้โดยวัดด้วยค่าของ AIC หรือค่าของ BIC
5. ประมาณค่าพารามิเตอร์ทั้งหมดโดยใช้วิธี Maximum likelihood estimation
6. สร้างแบบจำลองอนุกรมเวลา SARIMA ตามสมการที่ 2

งานวิจัยนี้สามารถพยากรณ์การจราจรของเครือข่ายไร้สายได้โดยมีค่าความผิดพลาดสัมพัทธ์ระหว่างค่าการจราจรที่ได้จากการพยากรณ์กับค่าการจราจรจริงมีค่าน้อยกว่า 0.02

บทที่ 3
หลักการของระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา

ในบทนี้จะกล่าวถึงสถาปัตยกรรม เครื่องมือที่ใช้ในการวิจัย การสร้างแบบจำลองอนุกรมเวลาซาริมา และการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา ดังนี้

3.1 สถาปัตยกรรม



ภาพที่ 3 สถาปัตยกรรม

ระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมามีองค์ประกอบ 3 ส่วน ดังนี้

1. หน่วยสังเกตการจราจรของเครือข่าย (Network traffic monitoring unit) มีหน้าที่คอยสังเกตการจราจรของเครือข่ายเพื่อวัดอัตราการใช้แบนด์วิดท์ที่ถูกใช้งานในเวลาต่างๆของเว็บเซิร์ฟเวอร์ทางการค้า โดยใช้ Zabbix [12] และส่งข้อมูลอัตราการใช้แบนด์วิดท์ไปให้กับหน่วยวิเคราะห์อัตราการใช้แบนด์วิดท์
2. หน่วยวิเคราะห์อัตราการใช้แบนด์วิดท์ (Bandwidth analyzing unit) มีหน้าที่วิเคราะห์อัตราการใช้แบนด์วิดท์ในเวลาปัจจุบันโดยเปรียบเทียบกับอัตราการใช้แบนด์วิดท์ที่ได้จากการพยากรณ์ที่ใช้แบบจำลองอนุกรมเวลาซาริมา ซึ่งสร้างตามอัลกอริทึมในงานวิจัย [9] ว่ามีความแตกต่างกันเกินค่าขีดสุดหรือไม่ ถ้าเกิดค่าขีดสุดจะถือว่ามีความผิดปกติเกิดขึ้น และจะส่งข้อมูลความผิดปกติไปให้กับหน่วยฐานข้อมูล
3. หน่วยฐานข้อมูล (Database Unit) มีหน้าที่จัดเก็บข้อมูลความผิดปกติที่เกิดขึ้นในเครือข่าย ณ เวลาต่างๆ

3.2 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยมีดังนี้

3.2.1 ภาษาอาร์ (R programming language)

ภาษาอาร์เป็นภาษาเขียนโปรแกรมสำหรับการประมวลผลทางสถิติและกราฟิก ภาษาอาร์เป็นภาษาที่มีการใช้งานอย่างแพร่หลายในหมู่นักสถิติสำหรับการพัฒนาซอฟต์แวร์ทางสถิติและวิเคราะห์ข้อมูล การวิจัยนี้นำภาษาอาร์มาใช้ในการสร้างแบบจำลองอนุกรมเวลาซาริมาและแบบจำลองอนุกรมเวลาอาริมา

3.2.2 GNU Plot

โปรแกรม GNU Plot คือ โปรแกรมสร้างกราฟสองมิติและสามมิติสำหรับฟังก์ชันและข้อมูลต่างๆ ซึ่งถูกเขียนด้วยภาษาซี การวิจัยนี้นำโปรแกรม GNU Plot มาใช้ในการสร้างกราฟสำหรับการจราจรของเครือข่าย

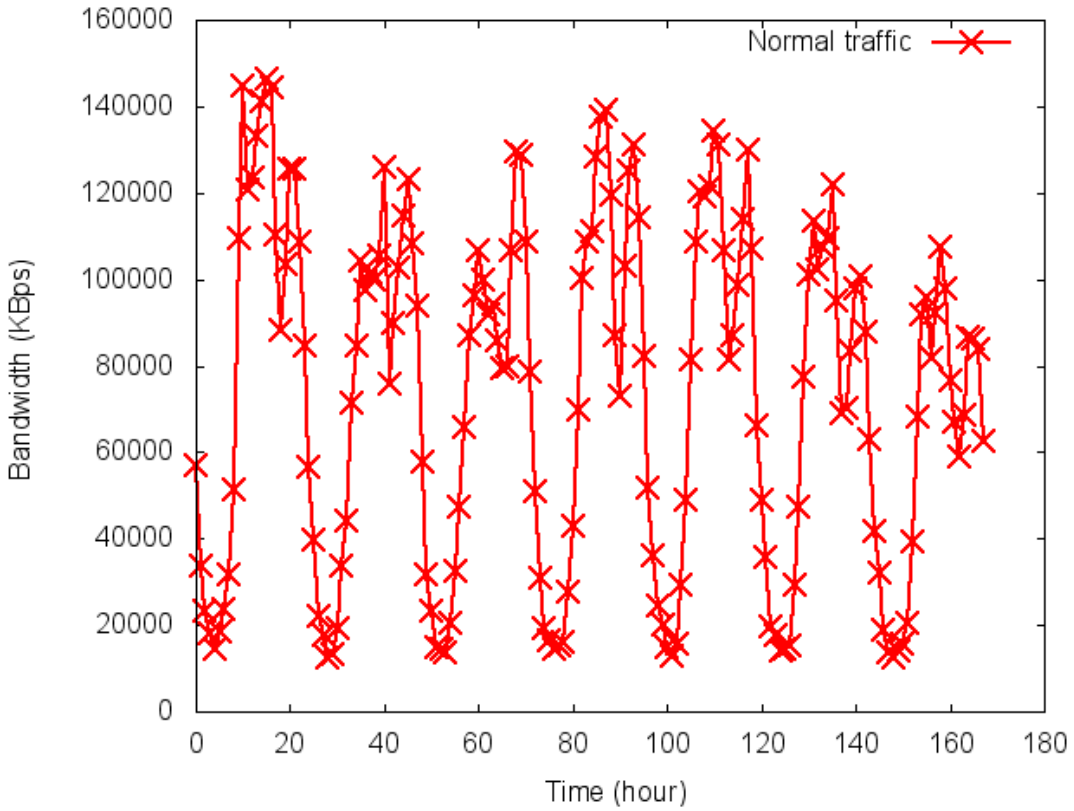
3.2.3 Zabbix

Zabbix คือ ระบบจัดการเครือข่าย ซึ่งถูกออกแบบมาเพื่อตรวจตราสถานะของบริการเครือข่าย ผู้ให้บริการ และฮาร์ดแวร์เครือข่ายต่างๆ ในการวิจัยนี้นำ Zabbix มาใช้ในการเก็บข้อมูลอัตราการใช้แบนด์วิดท์ของเครือข่าย

3.3 การสร้างแบบจำลองอนุกรมเวลาซาริมา

การวิจัยนี้นำข้อมูลอัตราการใช้แบนด์วิดท์ของเว็บเซิร์ฟเวอร์ทางการค้าที่ได้จาก Zabbix ตั้งแต่วันที่ 24 ธันวาคม 2553 ถึง 17 มีนาคม 2554 รวมทั้งหมด 2016 ชั่วโมงมาใช้ในการสร้างแบบจำลองอนุกรมเวลาซาริมา โดยใช้หนึ่งชั่วโมงเป็นหน่วยเวลา เพื่อลดการเกิดการเพิ่มขึ้นหรือลดลงอย่างรวดเร็ว (Spike) ในขณะที่ยังคงรูปร่างดั้งเดิมของอัตราการใช้แบนด์วิดท์ไว้ นอกจากนี้ยังมีการนำ BIC มาใช้ในการเลือกแบบจำลองที่มีความสูญเสียของข้อมูลน้อย อัลกอริทึมการสร้างแบบจำลองอนุกรมเวลาซาริมาถูกเสนอไว้ในงานวิจัย [9] ซึ่งการวิจัยนี้ดำเนินการตามขั้นตอนหลักดังนี้

1. หาค่าของคาบฤดูกาล s จากการวิเคราะห์กราฟ เมื่อวิเคราะห์การจราจรของเครือข่ายตั้งแต่วันที่ 24 ถึง 30 ธันวาคม 2553 ดังภาพที่ 4 จะได้ว่าค่าของคาบเท่ากับ 24 ชั่วโมง



ภาพที่ 4 การจราจรของเครือข่ายตั้งแต่วันที่ 24 ถึง 30 ธันวาคม 2553

2. หาค่าของ d และ D โดยการวิจัยนี้ใช้ค่าของ d เท่ากับ 1 เพราะการจราจรของเครือข่ายมีแนวโน้มเป็นเส้นตรง (Linear trend) และใช้ค่าของ D เท่ากับ 1 เพื่อให้แบบจำลองไม่สูญเสียรูปแบบเชิงฤดูกาล (Seasonal pattern) ในการพยากรณ์ระยะยาว (Long term forcast) [13]
3. หาค่าอันดับของ p q P และ Q โดยการทดสอบแทนค่า 0 1 และ 2 ทุกรูปแบบ ซึ่ง p และ P ไม่ควรเป็น 0 พร้อมกัน และ q และ Q ไม่ควรเป็น 0 พร้อมกัน จากนั้นเลือกรูปแบบการแทนค่าที่เหมาะสมที่สุดมาใช้โดยวัดด้วยค่าของ BIC [14] ในการวิจัยนี้เมื่อทดสอบการแทนค่าทุกรูปแบบได้ผลลัพธ์ว่าเมื่อค่าอันดับของ p q P และ Q เท่ากับ 0 1 2 และ 1 ตามลำดับจะได้ค่าของ BIC ดีที่สุดเท่ากับ 18.51
4. ประมาณค่าพารามิเตอร์ที่เหลือทั้งหมด เช่น ϕ_1 θ_1 Φ_1 และ Θ_1 โดยใช้วิธี Maximum likelihood estimation [15]
5. สร้างแบบจำลองอนุกรมเวลา SARIMA ตามสมการที่ 2

3.4 การตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา

การตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายสามารถทำได้โดยการเปรียบเทียบค่าของ NRMSE ของแต่ละวันกับค่าขีดสุด ถ้า NRMSE ของวันใดมีค่ามากกว่าค่าขีดสุด แสดงว่ามีความผิดปกติแบนด์วิดท์ของเครือข่ายเกิดขึ้น ซึ่งการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมามีขั้นตอนดังนี้

1. สร้างแบบจำลองอนุกรมเวลาซาริมาโดยใช้วิธีของงานวิจัย [9]
2. หาระยะเวลาที่เหมาะสมสำหรับการพยากรณ์จากข้อมูลที่ได้จาก Zabbix ระหว่างวันที่ 24 ธันวาคม 2553 ถึง 31 มีนาคม 2554 รวมทั้งหมด 2352 ชั่วโมง โดยเปรียบเทียบค่าของ BIC และเวลาที่ใช้ดำเนินการ (Execution time) ในการวิจัยนี้ทดสอบตั้งแต่การใช้ระยะเวลา 672 ชั่วโมง (4 สัปดาห์) ถึง 2352 ชั่วโมง (14 สัปดาห์) ดังตารางที่ 1

ตารางที่ 1 เปรียบเทียบ BIC และเวลาที่ใช้ดำเนินการ

ชั่วโมง	BIC	**เวลาที่ใช้ดำเนินการ (วินาที)**
672	18.69305	4
840	18.63797	5
1008	18.65222	6
1176	18.60795	7
1344	18.60625	8
1512	18.56333	9
1680	18.53564	10
1848	18.52005	14
2016	18.51008	17
2184	18.48406	19
2352	18.48369	21

3. จากตารางที่ 1 เปลี่ยนค่าของ BIC และเวลาที่ใช้ดำเนินการให้อยู่ในรูปของค่าความดีในการนำมาใช้พยากรณ์ ซึ่งค่าของ BIC ยิ่งน้อยยิ่งดี เนื่องจากมีความสูญหายของข้อมูลน้อย และเวลาที่ใช้ดำเนินการยิ่งน้อยยิ่งดี โดยนำตารางที่ 1 มา

เทียบบัญญัติไตรยางค์ และกำหนดให้ค่าความดีสูงสุดของค่าของ BIC และเวลาที่ใช้ดำเนินการเท่ากับ 50 เท่ากัน เพื่อให้ค่าความดีรวมสูงสุดมีค่าเท่ากับ 100 เนื่องจากต้องการให้ค่าของ BIC และเวลาที่ใช้ดำเนินการมีความสำคัญเท่ากัน เมื่อดูจากค่าความดีรวม ดังตารางที่ 2 ในการวิจัยนี้พบว่าระยะเวลาที่เหมาะสมสำหรับการพยากรณ์คือ 1680 ชั่วโมง จากนั้นจึงแบ่งข้อมูลจาก Zabbix ออกเป็นสองส่วนคือ ข้อมูลตั้งแต่วันที่ 24 ธันวาคม 2553 ถึง 3 มีนาคม 2554 รวมทั้งหมด 1680 ชั่วโมง เป็นข้อมูลสำหรับฝึก (Training data) และข้อมูลตั้งแต่วันที่ 4 มีนาคม 2554 ถึง 28 เมษายน 2554 เป็นข้อมูลสำหรับทดสอบ (Testing data)

ตารางที่ 2 ค่าความดี

ชั่วโมง	**ค่าความดีของ BIC**	**ค่าความดีของเวลาที่ใช้ดำเนินการ**	**ค่าความดีรวม**
672	0.00	50.00	50.00
840	13.15	47.06	60.21
1008	9.75	44.12	53.87
1176	20.32	41.18	61.50
1344	20.73	38.24	58.97
1512	30.98	35.29	66.27
1680	**37.59**	**32.35**	**69.95**
1848	41.32	20.59	61.90
2016	43.70	11.76	55.46
2184	49.91	5.88	55.79
2352	50.00	0.00	50.00

4. พยากรณ์อัตราการใช้แบนด์วิดท์ในชั่วโมงถัดไปในข้อมูลสำหรับทดสอบ โดยใช้ข้อมูล 1680 ชั่วโมงก่อนหน้าซึ่งในรอบแรกเป็นข้อมูลสำหรับฝึก ทำขั้นตอนนี้ซ้ำจนกระทั่งได้ค่าพยากรณ์อัตราการใช้แบนด์วิดท์ของทั้งวัน
5. หาผลต่างระหว่างค่าอัตราการใช้แบนด์วิดท์ที่พยากรณ์ได้กับค่าอัตราการใช้แบนด์วิดท์จริง

6. คำนวณค่า NRMSE ของทั้งวันโดยใช้สมการที่ 4 ซึ่งการวิจัยนี้ใช้ NRMSE ของทั้งวันเพื่อกำจัดการเพิ่มขึ้นหรือลดลงอย่างรวดเร็ว บางส่วนที่จะเกิดขึ้นถ้าใช้ค่า NRMSE ของแต่ละชั่วโมง
7. ทำขั้นตอนที่ 3 ถึง 5 ซ้ำจนกระทั่งได้ค่า NRMSE ของทั้งข้อมูลสำหรับทดสอบ
8. หาค่าขีดสุดจากขอบเขตบน (Upper bound) ของค่าของ NRMSE ทั้งหมด ด้วยการกำหนดให้ค่าเริ่มต้นของขอบเขตบนเป็น 0 แล้วเพิ่มค่าของขอบเขตบนครั้งละ 0.5 จนกระทั่งมีค่าที่มากกว่าขอบเขตบนน้อยกว่า 7.5 เปอร์เซนต์ของจำนวน NRMSE ทั้งหมด ซึ่งการวิจัยนี้ได้ค่าขีดสุดเท่ากับ 8.5 เปอร์เซนต์
9. นำค่าขีดสุดที่ได้ไปใช้งานจริง โดยเปรียบเทียบค่าของ NRMSE ของแต่ละวันกับค่าขีดสุด ถ้าค่าของ NRMSE มากกว่าค่าขีดสุดแสดงว่ามีความผิดปกติเกิดขึ้น หากต้องการทราบผลการตรวจหาความผิดปกติภายใน 1 ชั่วโมง ให้เปรียบเทียบค่าของ NRMSE ของ 24 ชั่วโมงก่อนหน้ากับค่าขีดสุดทุกรายชั่วโมงแทน

บทที่ 4
การทดสอบประสิทธิภาพของระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา

ในบทนี้จะกล่าวถึงหลักการของการทดสอบประสิทธิภาพ และผลลัพธ์ของการทดสอบประสิทธิภาพระบบตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา ดังนี้

4.1 หลักการของการทดสอบประสิทธิภาพ

การวิจัยนี้ทดสอบประสิทธิภาพของการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมาทั้งหมด 3 ส่วน โดยใช้การเปรียบเทียบ NRMSE กับค่าขีดสุดซึ่งมีค่าเท่ากับ 8.5 เปอร์เซนต์ ถ้าค่าของ NRMSE มีค่ามากกว่า 8.5 เปอร์เซนต์แสดงว่าตรวจหาความผิดปกติพบ แต่ถ้าค่าของ NRMSE มีค่าน้อยกว่า 8.5 เปอร์เซนต์แสดงว่าตรวจหาความผิดปกติไม่พบ

การทดสอบประสิทธิภาพส่วนแรกคือการทดสอบเพื่อหาแนวโน้มของปัจจัยต่างๆ ด้วยการจำลองการโจมตีเพื่อให้เกิดการหยุดการให้บริการในการจราจรของเครือข่ายวันที่ 29 มีนาคม 2554 ซึ่งมีปัจจัยที่ทดสอบมีทั้งหมด 3 ปัจจัยได้แก่ อัตราการใช้แบนด์วิดท์ ระยะเวลา และเวลาของการโจมตี

1. อัตราการใช้แบนด์วิดท์ทดสอบ 3 กรณีได้แก่
   - 10 เปอร์เซนต์ของอัตราการใช้แบนด์วิดท์สูงสุด
   - 25 เปอร์เซนต์ของอัตราการใช้แบนด์วิดท์สูงสุด
   - 50 เปอร์เซนต์ของอัตราการใช้แบนด์วิดท์สูงสุด
2. ระยะเวลาทดสอบ 3 กรณีได้แก่
   - 1 ชั่วโมง
   - 2 ชั่วโมง
   - 4 ชั่วโมง
3. เวลาของการโจมตีทดสอบ 2 กรณีได้แก่
   - 6.00 น.
   - 18.00 น.

การทดสอบการทดสอบประสิทธิภาพส่วนที่สองคือการทดสอบเพื่อหาผลลัพธ์ของการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายทุกรายชั่วโมง โดยทำการทดสอบจำลองการโจมตีที่

มีอัตราการใช้แบนด์วิดท์ 25 เปอร์เซนต์ของอัตราการใช้แบนด์วิดท์สูงสุด เป็นเวลา 2 ชั่วโมง ณ เวลา 18.00 น. ของวันที่ 29 มีนาคม 2554

การทดสอบประสิทธิภาพส่วนที่สามคือการทดสอบเพื่อหาค่าผลบวกลวงและผลลบลวง โดยการสุ่มจำลองการโจมตีที่มีอัตราการใช้แบนด์วิดท์ตั้งแต่ 5 เปอร์เซนต์ถึง 100 เปอร์เซนต์ เป็นเวลา 1 ชั่วโมง ตั้งแต่วันที่ 4 มีนาคม 2554 ถึง 28 เมษายน 2554 และวิเคราะห์ค่าของ NRMSE ของข้อมูลตั้งแต่วันที่ 4 มีนาคม 2554 ถึง 28 เมษายน 2554 แล้วสร้างเมตริกซ์ความสับสน

4.2 ผลลัพธ์ของการทดสอบประสิทธิภาพ

จากการทดสอบประสิทธิภาพส่วนแรกพบว่าค่าของ NRMSE ขึ้นอยู่กับอัตราการใช้แบนด์วิดท์ของการโจมตี ถ้าอัตราการใช้แบนด์วิดท์มาก ค่าของ NRMSE จะมีค่ามาก แต่ถ้าอัตราการใช้แบนด์วิดท์น้อย ค่าของ NRMSE จะมีค่าน้อย นอกจากนั้นผลของเวลาของการโจมตีเป็นแบบสุ่ม ผลลัพธ์ของการทดสอบประสิทธิภาพถูกแสดงไว้ดังตารางที่ 3 และสามารถสรุปได้ว่าวิธีนี้สามารถตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายได้อย่างมีประสิทธิภาพ อย่างไรก็ตามวิธีนี้ไม่สามารถตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายที่มีอัตราการใช้แบนด์วิดท์น้อย และความผิดปกติแบนด์วิดท์ที่เกิดขึ้นต่อเนื่องเป็นเวลานานได้ เนื่องจากแบบจำลองอนุกรมเวลาจะปรับตัวเข้ากับความผิดปกติแบนด์วิดท์

ตารางที่ 3 ผลลัพธ์ของการทดสอบประสิทธิภาพเพื่อหาแนวโน้มของปัจจัยต่างๆ

กรณีที่	แบนด์วิดท์	ระยะเวลา (ชั่วโมง)	เวลา	NRMSE	ผลลัพธ์
1	10%	1	6.00 น.	6.92	ไม่พบ
2	10%	1	18.00 น.	7.98	ไม่พบ
3	10%	2	6.00 น.	6.54	ไม่พบ
4	10%	2	18.00 น.	7.44	ไม่พบ
5	10%	4	6.00 น.	5.82	ไม่พบ
6	10%	4	18.00 น.	6.54	ไม่พบ
7	25%	1	6.00 น.	9.3	พบ
8	25%	1	18.00 น.	9.98	พบ
9	25%	2	6.00 น.	8.85	พบ
10	25%	2	18.00 น.	9.31	พบ
11	25%	4	6.00 น.	8.05	ไม่พบ

12	25%	4	18.00 น.	8.20	ไม่พบ
13	50%	1	6.00 น.	15.11	พบ
14	50%	1	18.00 น.	12.5	พบ
15	50%	2	6.00 น.	15.03	พบ
16	50%	2	18.00 น.	12.12	พบ
17	50%	4	6.00 น.	14.23	พบ
18	50%	4	18.00 น.	11.42	พบ

จากการทดสอบประสิทธิภาพส่วนที่สองได้ผลลัพธ์การตรวจหาความผิดปกติแบนด์วิดท์ ดังตารางที่ 4

ตารางที่ 4 ผลลัพธ์การตรวจหาความผิดปกติแบนด์วิดท์ทุกรายชั่วโมง

ชั่วโมง	NRMSE	**ผลลัพธ์**
1-24	5.781095	ไม่พบ
2-25	5.942751	ไม่พบ
3-26	6.047077	ไม่พบ
4-27	6.127328	ไม่พบ
5-28	6.11308	ไม่พบ
6-29	6.140169	ไม่พบ
7-30	6.155006	ไม่พบ
8-31	6.282418	ไม่พบ
9-32	6.193404	ไม่พบ
10-33	6.170716	ไม่พบ
11-34	5.787008	ไม่พบ
12-35	6.777815	ไม่พบ
13-36	5.588905	ไม่พบ
14-37	5.614263	ไม่พบ
15-38	5.837235	ไม่พบ
16-39	5.663106	ไม่พบ

17-40	5.725023	ไม่พบ
18-41	5.675729	ไม่พบ
19-42	5.830783	ไม่พบ
20-43	8.675166	**พบ**
21-44	8.528687	**พบ**
22-45	9.085651	**พบ**
23-46	9.395884	**พบ**
24-47	9.428176	**พบ**
25-48	9.515454	**พบ**
26-49	9.494159	**พบ**
27-50	9.437698	**พบ**
28-51	9.386579	**พบ**
29-52	9.388624	**พบ**
30-53	9.392019	**พบ**
31-54	9.432827	**พบ**
32-55	9.401386	**พบ**
33-56	9.312704	**พบ**
34-57	9.305542	**พบ**
35-58	9.325447	**พบ**
36-59	8.740581	**พบ**
37-60	10.27821	**พบ**
38-61	10.63626	**พบ**
39-62	10.54916	**พบ**
40-63	10.55814	**พบ**
41-64	10.55364	**พบ**
42-65	10.52431	**พบ**
43-66	10.53967	**พบ**
44-67	9.418829	พบ
45-68	9.678167	พบ
46-69	8.567426	พบ
47-70	7.96793	ไม่พบ

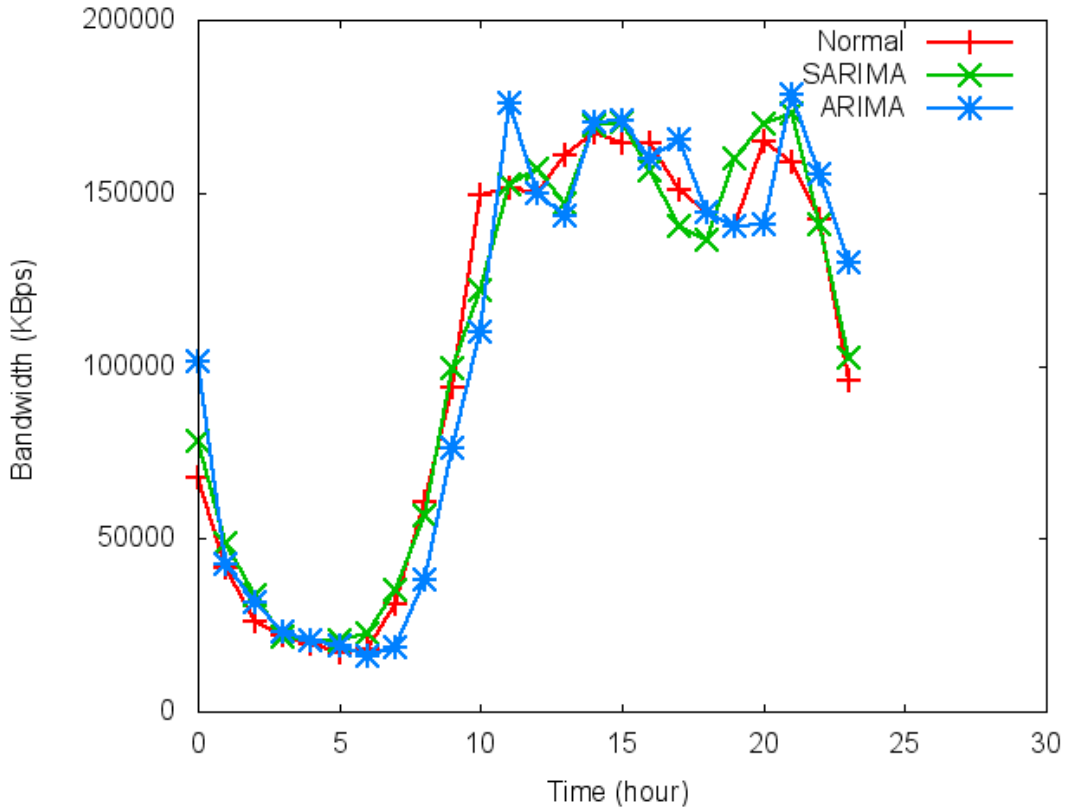
48-71	7.9446	ไม่พบ
49-72	7.804152	ไม่พบ

จากการทดสอบประสิทธิภาพส่วนที่สามสามารถสร้างเมตริกซ์ความสับสนได้ดังตารางที่ 5 เมื่อนำมาคำนวณหาอัตราผลบวกลวงจากจำนวนผลบวกลวงหารด้วยจำนวนผลบวกลวงบวกกับจำนวนผลลบจริง และคำนวณหาอัตราผลลบลวงจากจำนวนผลลบลวงหารด้วยจำนวนผลลบลวงบวกกับจำนวนผลบวกจริง จะได้ว่ามีอัตราผลบวกลวงเท่ากับ 3.57 เปอร์เซนต์ และอัตราผลลบลวงเท่ากับ 26 เปอร์เซนต์

ตารางที่ 5 เมตริกซ์ความสับสน

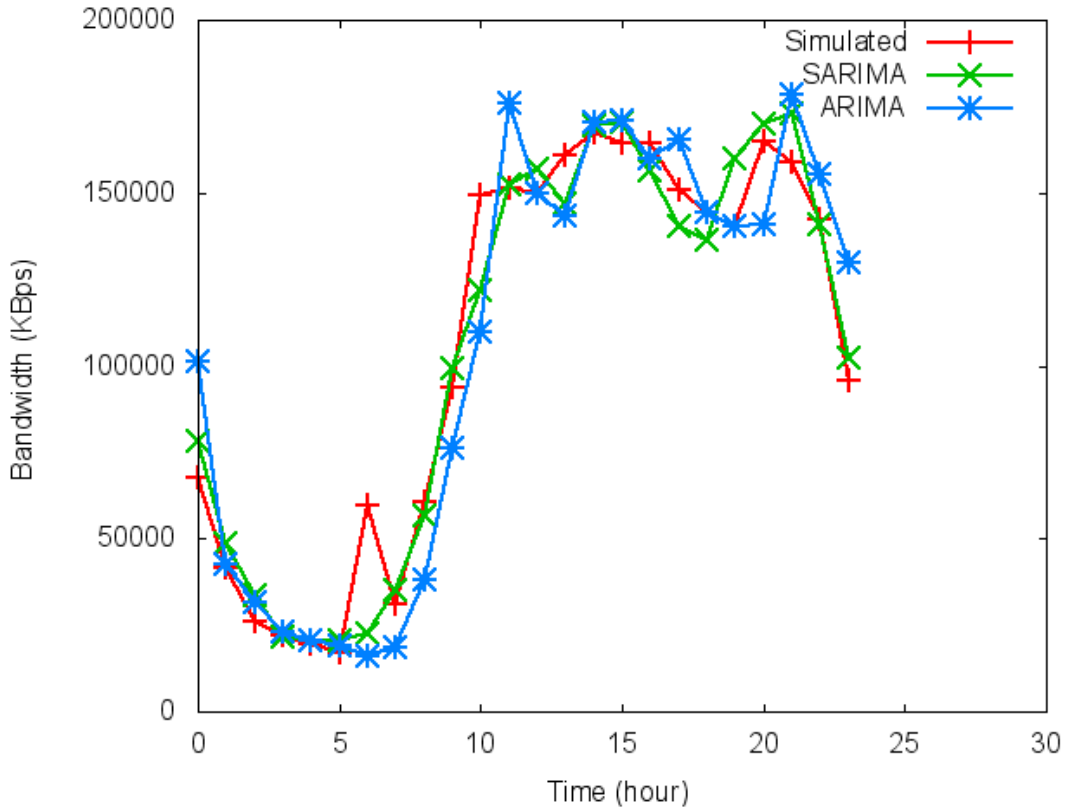
		ค่าที่ได้จากการพยากรณ์	
		ปกติ	**ผิดปกติ**
ค่าจริง	**ปกติ**	54	2
	ผิดปกติ	26	74

ค่าอัตราการใช้แบนด์วิดท์ของเครือข่ายที่ได้จากการพยากรณ์โดยใช้แบบจำลองอนุกรมเวลาซาริมาและแบบจำลองอนุกรมเวลาอาริมาค่อนข้างใกล้เคียงกับค่าอัตราการใช้แบนด์วิดท์จริงของเครือข่าย แต่แบบจำลองอนุกรมเวลาอาริมาจะมีค่าความผิดพลาดที่สูงกว่า เช่น การจราจรปกติของเครือข่ายวันที่ 29 มีนาคม 2554 ดังภาพที่ 5 จะพบว่าแบบจำลองอนุกรมเวลาอาริมามีค่าความผิดพลาดสูงที่ชั่วโมงที่ 10 11 และ 23 ในขณะที่แบบจำลองอนุกรมเวลาซาริมามีค่าความผิดพลาดที่ต่ำกว่ามาก



ภาพที่ 5 การจราจรปกติของเครือข่ายวันที่ 29 มีนาคม 2554

เมื่อพิจารณาถึงการตรวจหาความผิดปกติแบนด์วิดท์จะพบว่าแบบจำลองอนุกรมเวลาซาริมาให้ผลลัพธ์ที่ดีกว่าแบบจำลองอนุกรมเวลาอาริมา เพราะสามารถพยากรณ์อัตราการใช้แบนด์วิดท์ของเครือข่ายได้ใกล้เคียงอัตราการใช้แบนด์วิดท์จริงของเครือข่ายมากกว่า เมื่อนำการจราจรปกติของเครือข่ายวันที่ 29 มีนาคม 2554 มาจำลองการโจมตีที่มีอัตราการใช้แบนด์วิดท์เท่ากับ 25 เปอร์เซนต์ของอัตราการใช้แบนด์วิดท์สูงสุดในชั่วโมงที่ 6 ดังภาพที่ 6 จะเห็นได้ว่าแบบจำลองอนุกรมเวลาซาริมาสามารถแยกการโจมตีออกจากการจราจรปกติได้ ในขณะที่แบบจำลองอนุกรมเวลาอาริมาไม่สามารถแยกการโจมตีออกจากการจราจรปกติในชั่วโมงที่ 10 และ 11 ได้



ภาพที่ 6 การจราจรของเครือข่ายวันที่ 29 มีนาคม 2554 โดยมีการจำลองการโจมตี

จากการวิเคราะห์ผลลัพธ์ของการทดสอบประสิทธิภาพพบว่าการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาซาริมามีข้อดีข้อเสียโดยรวมดังนี้

1. ข้อดี
   - ไม่ขึ้นอยู่กับเวลาของการโจมตี เนื่องจากผลของเวลาของการโจมตีเป็นแบบสุ่ม
   - สามารถตรวจหาความผิดปกติแบนด์วิดท์ได้ทั้งอัตราการใช้แบนด์วิดท์ที่เพิ่มขึ้นอย่างผิดปกติ เช่น การโจมตี เป็นต้น และอัตราการใช้แบนด์วิดท์ที่ลดลงอย่างผิดปกติ เช่น การล้มเหลวของระบบ เป็นต้น เนื่องจากการคำนวณใช้ค่า NRMSE ซึ่งเป็นค่าสัมบูรณ์ (Absolute value)
   - สามารถตรวจหาความผิดปกติแบนด์วิดท์ได้แม่นยำกว่าแบบจำลองอนุกรมเวลาอาริมา เนื่องจากแบบจำลองอนุกรมเวลาซาริมาใช้ค่าขีดสุดที่ต่ำกว่าแบบจำลองอนุกรมเวลาอาริมา

- มีอัตราผลบวกลวงเพียง 3.57 เปอร์เซนต์ ซึ่งน้อยกว่าแบบจำลองอนุกรมเวลาอาริมาที่มีอัตราผลบวกลวง 5.36 เปอร์เซนต์

2. ข้อเสีย
   - ไม่สามารถตรวจหาความผิดปกติแบนด์วิดท์ที่มีอัตราการใช้แบนด์วิดท์ต่ำได้
   - ไม่สามารถตรวจหาความผิดปกติแบนด์วิดท์ที่เกิดขึ้นต่อเนื่องเป็นเวลานานได้

บทที่ 5
สรุปผลการวิจัย

ในบทนี้จะกล่าวถึงสิ่งที่ได้จากการวิจัย (Contribution) ประโยชน์ของการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา ข้อเสนอแนะ และสรุปผลการวิจัย ดังนี้

5.1 สิ่งที่ได้จากการวิจัย

สิ่งที่ได้จากงานวิจัยนั้นมีดังนี้

1. อธิบายปัญหาด้านความปลอดภัยของระบบเครือข่าย
2. นำเสนอวิธีการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาซาริมา
3. เปรียบเทียบความแตกต่างระหว่างการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาซาริมาและการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองอนุกรมเวลาอาริมา

5.2 ประโยชน์ของการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมา

การตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้ซาริมาสามารถตรวจหาความผิดปกติแบนด์วิดท์ได้แม่นยำกว่าการตรวจหาความผิดปกติของเครือข่ายโดยใช้อาริมา โดยเกิดผลบวกลวงน้อยกว่าและใช้ค่าขีดสุดต่ำกว่า

5.3 ข้อเสนอแนะ

การตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายมีแนวทางการพัฒนาต่อ ดังนี้

1. ออกแบบการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองฟาริมา (FARIMA) และเปรียบเทียบกับการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้แบบจำลองซาริมา
2. นำผลการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายโดยใช้อาริมา ซาริมา และฟาริมา มาผสมผสานกันเพื่อให้ได้ผลการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายที่แม่นยำมากขึ้น

5.4 สรุปผลการวิจัย

การวิจัยนี้เสนอวิธีการตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่าย โดยใช้ซาริมาในการพยากรณ์อัตราการใช้แบนด์วิดท์ของเครือข่าย เพื่อนำมาเปรียบเทียบกับอัตราการใช้แบนด์วิดท์จริงของเครือข่ายที่ได้จาก Zabbix ถ้าผลต่างระหว่างอัตราการใช้แบนด์วิดท์ของเครือข่ายทั้งสองมากกว่าค่าขีดสุดจะถือว่ามีความผิดปกติเกิดขึ้น จากการทดสอบประสิทธิภาพพบว่าวิธีนี้สามารถตรวจจับความผิดปกติแบนด์วิดท์ของเครือข่ายได้อย่างมีประสิทธิภาพ โดยใช้ค่าขีดสุดเท่ากับ 8.5 เปอร์เซนต์ของอัตราการใช้แบนด์วิดท์สูงสุดของแต่ละวัน และมีอัตราผลบวกลวงเท่ากับ 3.57 เปอร์เซนต์ และอัตราผลลบลวงเท่ากับ 26 เปอร์เซนต์

ข้อจำกัดของการวิจัยนี้คือ วิธีนี้ไม่สามารถตรวจหาความผิดปกติแบนด์วิดท์ของเครือข่ายที่มีอัตราการใช้แบนด์วิดท์ต่ำ และความผิดปกติแบนด์วิดท์ที่เกิดขึ้นต่อเนื่องเป็นเวลานานได้ อย่างไรก็ตามข้อจำกัดแรกสามารถละทิ้งได้เพราะมีผลกระทบเพียงเล็กน้อย และข้อจำกัดที่สองต้องการข้อมูลเพิ่มเติมนอกจากอัตราการใช้แบนด์วิดท์ เพื่อแยกแยะความผิดปกติแบนด์วิดท์ที่เกิดขึ้นต่อเนื่องเป็นเวลานานออกจากการปรับปรุงเครือข่าย

รายการอ้างอิง

[1] Shabtai, A., Fledel, Y., Kanonov, U., Elovici, Y., Dolev, S., and Glezer, C. Google android: A comprehensive security assessment. IEEE Security & Privacy Mag. 8, 2 (2010) : 35–44.

[2] Twitte. Twitter Status [Online]. Available from : http://status.twitter.com/post/157191978/ongoing-denial-of-service-attack [2010, September 1]

[3] Whitman, M. E., and Mattord, H. J. Principles of Information Security. Course Technology, 2008.

[4] Yaacob, A. H., Tan, I. K. T., Chien, S. F., and Tan, H. K. ARIMA Based Network Anomaly Detection. 2nd Int. Conf. on Communication Software and Networks (2010) : 205-209.

[5] Basu, S., Mukherjee, A., and Klivansky, S. Time Series Models for Internet Traffic. Proc. IEEE INFOCOM Conf. (1996) : 611-620.

[6] Hanbanchong, A., and Piromsopa, K. SARIMA Based Network Bandwidth Anomaly Detection. The Ninth International Joint Conference on Computer Science and Software Engineering (2012).

[7] Box, G., Jenkins, G., and Reinsel, C. Time Series Analysis: Forecasting and Control. Englewood Cliffs, NJ : Prentice-Hall, 1994.

[8] Hill, T., and Lewicki, P. STATISTICS: Methods and Applications. Tulsa : StatSoft, 2007.

[9] Shu, Y., Yu, M., Liu, J., and Yang, O. W. W. Wireless Traffic Modeling and Prediction Using Seasonal ARIMA Models. Proc. IEEE ICC Conf. 3 (2003) : 1675–1679.

[10] Burnham, A. Model Selection and Inference - A practical information-theoretic approach. 1998.

[11] Pfanzagl, J., and Hamböker, R. Parametric Statistical Theory. Berlin : Walter de Gruyter, 1994.

[12] Zabbix. Homepage of Zabbix [Online]. Available from : http://www.zabbix.com [2012, February 15]

[13] Nau., R. F. Seasonal ARIMA models [Online]. Available from : http://www.duke.edu/~rnau/seasarim.htm [2012, February 13]

[14] Akaike, H. Applied Time Series Analysis. New York : Academic Press, 1978.

[15] Haslett, J., and Raftery, A. E. Space-time modeling with long-memory dependence: assessing Ireland's wind power resource. Applied Statistics 38, 1 (1989) : 1-50.

ประวัติผู้เขียนวิทยานิพนธ์

นายอภิชิต หาญบรรจง เกิดเมื่อวันที่ 24 มกราคม พ.ศ.2531 ที่จังหวัดกรุงเทพฯ สำเร็จการศึกษาหลักสูตรวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ จากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2552 และเข้าศึกษาต่อในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2553