ให้ Squid เป็น URL หลัง "?"

ทริคเล็กๆ ที่หายากมาก เลยเอามาบอกต่อ

ระบบไฟร์วอลการเข้าเว็บบนลีนุกซ์ หรือที่รู้จักกันในชื่อ "Squid" นอกจากใช้ตั้ง Access-List ควบคุมข้อมูลเข้าออกได้แล้ว ยังใช้เก็บ Log (หรือเรียกว่า Cache) การเข้าเว็บต่างๆ จากเครื่องลูกในแลนได้ด้วย โดยไฟล์ Log ดิบที่อัพเดตตลอดเวลาอยู่ที่ /var/log/squid/access.log

ความรู้เพิ่มเติม: โปรแกรมไฟร์วอลหรือพ็อกซีบนลีนุกส์ที่ใช้ IPtable ฟอร์เวิร์ดพอร์ตไปใช้บริการนั้นมีหลายตัว แต่ละตัวจะจำเพาะกับแต่ละโปรโตคอล เช่น Squid จะใช้กับโปรโตคอล HTTP (พอร์ต 80) เช่น การเข้าเว็บไซต์ ส่วน Frox ใช้กับโปรโตคอล FTP (พอร์ต 21) เช่น การโหลดไฟล์แพชต่างๆ

แต่บังเอิญที่ ดีฟอลต์ของ Squid จะไม่เก็บข้อมูล URL ที่คิดว่าไม่จำเป็นและเปลืองเนื้อที่ รวมไปถึงข้อมูลหลัง "?" ด้วย ซึ่งปัจจุบันข้อมูล Query ตัวแปรส่วนนี้มีความสำคัญมาก เช่น ถ้าเก็บเฉพาะ "youtube.com/?" ก็ไม่รู้ว่ากำลังดูคลิปอะไรอยู่เป็นต้น (แล้วยิ่งการเก็บ Log ตาม พ.ร.บ.ฯ ต้องการข้อมูลละเอียดเสียด้วย)

ที่เรารู้ตอนแรกคือ ต้องเกี่ยวกับการตั้งค่าในไฟล์คอนฟิก /etc/squid/squid.conf แน่นอน แต่เราไม่รู้คำสั่งที่ใช้ โดยเฉพาะพารามิเตอร์นี้ไม่ได้มีเครื่องหมายคอมเม้นต์ ("#") พรางแสดงไว้ให้แก้ด้วยสิ ถามอากู๋หน้าแรกๆ ก็มีแต่ให้ดูที่ acl QUERY/cache QUERY ซึ่งแก้เท่าไรก็ไม่ได้ผล

ค้นไปค้นมาจึงเจอคำสั่งทีเด็ด นั่นคือ "strip_query_terms off" ไม่รอเช้า รีบใช้ WinSCP ดึงไฟล์ squid.conf มาเพิ่มคำสั่งนี้ (แทรกบรรทัดไหนก็ได้) จับคัดลอกเข้าไปใหม่ พร้อมสั่ง "service squid restart" บนเชลล์คอนโซล โอเค ไม่เจอข้อความ Error ว่าไม่รู้จักคำสั่งนี้ก็น่าจะเรียบร้อย

เรามาดูผลงานในไฟล์ acess.log บรรทัดล่าสุด ดังนี้

คำสั่ง tail ใช้ดูข้อมูลล่าสุด (หางไฟล์) ส่วนพารามิเตอร์ “-f” ใช้เพื่อติดตามความเคลื่อนไหวตลอด (Follow)

ใช้เครื่องหมาย “ | ” เพื่อส่งต่อ Output มาให้คำสั่งด้านหลัง (Pipeline) ส่วนคำสั่ง grep ใช้กรองข้อมูลเฉพาะบรรทัดที่มีข้อความที่ต้องการ

```
[root@system ~]# tail -f /var/log/squid/access.log | grep 192.168.1.247
2010-08-12 09:04:39    563 192.168.1.247 TCP_MISS/200 471 GET http://www.linuxthai.org
/forum/index.php? - DIRECT/210.1.61.48 text/html
2010-08-12 09:05:35    280 192.168.1.247 TCP_MISS/200 9190 GET
http://edge5.catalog.video.msn.com/videoByMarket.aspx? - DIRECT/65.54.82.152 text/xml
2010-08-12 09:07:22    314 192.168.1.247 TCP_MISS/304 303 GET
http://img1.catalog.video.msn.com/image.aspx?uuid=9a3defca-c81c-426b-8364-
fa70059087b0&w=136&h=102 - DIRECT/65.54.82.154 image/jpeg
```

Cache ตอนยังไม่ตั้งค่า “strip_query_terms off”

Cache หลังเปลี่ยนค่าใน squid.conf แล้วรีสตาร์ทเซอร์วิสแล้ว

เท่านี้ เราก็เก็บ URL ได้เต็มๆ (ถ้าไม่กลัวเนื้อที่เต็ม) ทีนี้ก็เหลือแต่มาทำ Logrotate เขียนสคริปต์หั่นไฟล์ล็อกนี้ให้เหลือเฉพาะข้อมูลที่ต้องการ (เช่น เอาเฉพาะคำที่ 1, 2, 4, 8, 9 ของแต่ละบรรทัดด้วยคำสั่ง | awk ‘{print $1 $2 $4 $8 $9}’ หรือต่อด้วยคำสั่ง | sort –r เพื่อให้เรียงกลับกัน (Reverse) จากข้อมูลบรรทัดล่าสุดลงไป เป็นต้น)

ที่เหลือก็ลองรันโค้ดหน้าเว็บ PHP รับค่าพวกนี้มาใช้งาน ทำอินเทอร์เฟซสวยๆ แล้วจับฮาร์ดแวร์ทั้งหมดลงบ็อกซ์เล็กๆ ไม่แน่ว่าท่านอาจจะสร้างตัวเก็บ Log ที่ขายดีกว่าพวกแบรนด์เนมอย่าง BlueCoat หรือ CS MARS ก็ได้