DECLARATION OF KEITH J. TERUYA

I, KEITH J. TERUYA, declare the following:

1. I have been retained by counsel for NetNut Ltd. to provide my opinions on certain issues in connection with revised substitute claims 109–125 of U.S. Patent No. 9,241,044 (the '044 patent), as shown in Ex. 2029 (Appendix A-2) in IPR2021-00458 ('044 Revised Amendments), and claims 29-38 of U.S. Patent No. 9,742,866 (the '866 patent), as shown in Ex. 2029 (Appendix A-2) in IPR2021-00465 ('866 Revised Amendments).

2. I have submitted declarations in each of the above-mentioned cases, in support of the Petitions therein, and in opposition to prior motions to amend made in each case. The facts concerning my compensation, background and qualifications, understanding of pertinent legal standards I have been asked to apply, and level of skill in the art are set forth in those declarations.

3. In further preparing my opinions as set forth herein, I have reviewed the '044 and '866 Revised Amendments, as well reviewed and/or re-reviewed the '044 and '866 patents, the original disclosures filed for them (Exs. 2022-2027 in each case, as well as Ex. 2028 in each case, showing where the patent owner points to for claim support), U.S. patent no. 10,469,614 (another patent owned by the present patent owner ("'614 patent")), all of the materials I previously stated in the earlier declarations that I had reviewed, the reexamination request, grant, and first

office action in case 90/014,880 concerning the ’614 patent, and references including Exs. 1017 (Shribman), 1105 (Reed (Onion Routing)), 1208 (Chauhan, Citrix), 1209 (Liu, Oracle), 1210 (Sigurdsson, Google), 1211 (Dorso), 1212 (Mithyantha, Citrix), 1213 (Jackowski, Citrix), 1214 (MorphMix, 2004 thesis by Marc Rennhard), and 1217/1218 (VIP72 materials published on YouTube).

4. My understanding of the ’044 and ’866 patents is that they share a common disclosure, and disclose a system of Internet content retrieval through distributed agent, peer, and tunnel devices, in which a tunnel device may be assigned to perform a retrieval task by facilities including an “acceleration server.” Each tunnel device monitors and via a heartbeat mechanism reports parameters concerning its own operating condition (such as CPU utilization and/or idle state, among other information) to the acceleration server, which maintains records of such information. When assigning tunnels to perform retrievals for requesting clients, the acceleration server can take into account the resource utilization on the tunnels, avoiding the selection of tunnels for which the utilization is above or below certain thresholds (*e.g.*, “congested”). This is reflected in the state diagram shown in Fig. 30 and the discussion in the specification concerning that figure *(e.g.*, around cols. 106-108 and 120-121 and the corresponding pages (*e.g.*, pp. 140, 163-165) of the originally filed disclosures, and how the resulting states are reported and used by the acceleration server.

5. The ’044 and ’866 patents also disclose piece-wise retrieval of content, and particularly that the agents and peers partition into “content slices” and store content they receive “fresh” from origin servers, so as to cache the content for later retrieval (including piece-wise retrieval by slice) by other clients (which reassemble them for presentation), where it may be advantageous, for example, for performance or availability, for the client to do so. *E.g.*, cols. 102-103 of the specification and corresponding pages of the original disclosures. However, the revised amended claims recite the condition “wherein the first tunnel device does not cache a copy of the first fresh content.” I could find nothing in the materials pointed to by the patent owner in Ex. 2028 (in the respective ’044 and ’866 cases), or in the disclosures themselves, that reflects a non-caching tunnel device.

6. The claims also recite the use of a tunneling protocol, *e.g.*, for a requesting client to send a content request to a tunnel device. In general, it is very difficult for a tunnel to cache information passed through it via a tunneling protocol, as it acts as a pass-through in that case, and doesn’t necessarily “see” the contents or its framing (when any unit of information actually begins or ends). In any case, however, the revised amended claims only recite the use of a tunneling protocol to get the content request to the tunnel device, and not that the content fetch to and from the content server is carried out over a tunnelling protocol. This

therefore has no bearing on whether the tunnels would cache retrieved content, and is most consistent with caching operation, which is what is repeatedly disclosed. Perhaps this system could have been constructed in some way without caching by the tunnels, but I do not see in these disclosures how that would or could be done, or even that such a mode of operation was being disclosed at all. Further, in my opinion, the claims of the ’866 patent in particular—as best I can understand them because in some respects they are far from precise, as well as self-contradictory—would not be operable without the ability of the tunnel device to cache the fresh content that it received.

7. The revised amended claims now introduce a feature, very much resembling what is already in the patent owner’s ’614 patent, whereby a tunnel is assigned (and an identifier for it received by a client, and the tunnel is used to receive content requests), “so long as” its “idle resource” does not “meet the criterion of being “above or below a set threshold.”

8. Work load and balancing it is a concern when distributing computing tasks to a plurality of devices, particular those that have other workloads. Shribman (Ex. 1017) presents such concerns, because its distributed devices, as in the ’044/’866 patents, are consumer devices. This factor of using devices that have other tasks and responsibilities that their users may care about would have been a powerful motivation for a POSITA designing a system similar in object and

functionality to Shribman, which likewise relied on consumer devices to perform computing tasks. Shribman already disclosed a highly analogous checking operation for tunnels, to the extent of having the acceleration server check the online or offline state of the device before assigning it. Adding an additional factor of device congestion, determined by whether CPU usage (or conversely idling) was above or below a set threshold, would have avoided further burdening congested devices, and avoided relying on congested devices, which would work to improve the experience of the user base as a whole, as well as to more efficient operations.

9. A POSITA would have realized that such load-based assignment was readily achievable in the context of Shribman because analogous practices were well-known and commonplace. A POSITA could have observed others doing the same, in systems such as those described by Dorso (Ex. 1211), Liu (Ex. 1209), Chauhan (Ex. 1208), MorphMix (Ex. 1214), Siggurdsson (Ex. 1210), Mithyantha (Ex. 1112), and Jackowski (Ex. 1213), and many others. Moreover, Shribman with its selection based on online/offline state, already provides an opportune hook on which to place such a feature, simply by adding a "congested" state to the existing possible online-offline states, and arriving at what is now claimed. It does not matter whether the suggestion is nominally based on CPU utilization or "idling"—the one is the flip side (inverse) of the other, with idling being even more broad

than that, as it can include other things, such as idle network connections. This equivalence is recognized in the '044/'866 patents themselves. Ex. 1001 at 118:27-28 ("CPU utilization relates to the relative time that the CPU is not idling"), and 154:65-155:9 (also relating "idleness resource" to "idleness of communication."

10. For example, Dorso's peer-to-peer content retrieval environment includes relay peers that reasonably correspond in role and function to the tunnel devices of the '044/'866 patents. Dorso at ¶¶ 0050, 0059, discloses "optimizing" the selection of relay pees based on "factors related to the capabilities and resources, such as CPU and bandwidth." Chauhan does likewise, using "load balancing virtual servers" in distributed network appliances for functions such as "expir[ing] client connections by tracking the amount of time a connection has been idle," (¶ 0240), or "terminating a virtual machine" (¶ 0240), and that back-end servers may be selected by a load balancing virtual server "that manages the load placed on each server" (¶ 0294). These things are what a POSITA would need to know to incorporate a criterion based on resource utilization of idling into Shribman to arrive at the present claim. The techniques taught in Dorso and/or Chauhan would have been readily applicable to Shribman, with predictable results.

11. Liu, Ex. 1210, addresses management of network operations by distributed agents that monitor the state of distributed servers. At ¶ 0037, Liu discloses agents that collect data on "resource usage, performance, workload,

etc.,'" and "transform or classify collected data … as discrete values (*e.g.*, CPU usage equals 'high,' 'medium,' or 'low'), according to programmable boundaries." This fully corresponds to the claimed comparison with a "set threshold." Liu also tracks device state with a "state machine" (*see* Fig. 5), where state transitions reflect a measured condition crossing such a set threshold.

12. Mithyantha, Ex. 1212, discloses "status monitor 704 may determine that a monitored status or value exceeds a threshold, such as a CPU utilization rate past a threshold for a predetermined time," causing it to be removed from routing (so that it will no longer receive requests). 57:11-14, 2:15-19, 3:4-8, claim 4.

13. Jackowski, Ex. 1213, discloses at ¶ 0220 a connection manager that can manage a list of peers and detect congestion (corresponding to idleness of its connection below a threshold), and in ¶ 0221 sending a request to a peer via a tunnel to establish a new connection after the connection manager determines that the new peer connection has not been flagged for congestion.

14. MorphMix, Ex. 1214, addresses conditions under which a content relaying node in a peer-based retrieval network may or may accept a role in retrieving content, including whether it has "spare resources'" (p. 134), such as "computing power" (p. 146).

15. Sigurdsson, Ex. 1210, discloses a peer-based editing system that teaches (at ¶ 0061) determining whether a peer is available is performed by the client device "monitoring polling by the client."

16. Furthermore, these references also teach the claimed detail of determining a device assignment criterion by comparing resource utilization or idling to determine if it is above or below a set threshold. For example:

17. Mithyantha: "status monitor 704 may determine that a monitored status or value exceeds a threshold, such as a CPU utilization rate past a threshold for a predetermined time." Ex.\ 1212 at 57:11-14, 2:15-19, 3:4-8, claim 4.

18. Liu: ¶ 0058: "As illustrated in FIG. 4, if the data collection agent detects a change of state (e.g., a transition from 'medium high' CPU usage to 'high' CPU usage), a threshold value that is reached, or a threshold rate of change that is reached … it may send a message to the information layer agent indicating this event"---which is substantially identical to the corresponding disclosure of this feature in the '044 and '866 patents.

19. Dorso: ¶¶ 0031-32, 0050. Given that Dorso teaches to use an algorithm to optimize relay peer selection based on CPU resources, a POSITA armed with that knowledge would know how to implement that teaching by the simple method of compare observed levels of CPU activity (or conversely, idling) to a set threshold, in order to make a selection on that basis.

20. Furthermore, comparison with a threshold and triggering an action based on the comparison is a stock computing technique from the dawn of computing, for making anything that can be measured through computer I/O become actionable. Though not the earliest example by any means, see for example, Ex. 1222, Fry, *et al.*, U.S. Pat. No. 4,403,286 A (IBM, filed 1981), abstract, 6:21-48 (when "busy state" "tally" exceeding a threshold triggers a to reallocate the workload). Such techniques would be second nature to a POSITA in order to implement the very solutions described in any of Chauhan, MorphMix (determining "spare resources"), Sigurdsson, or Jackowski, without having to be expressly told to "compare it to a set threshold"—although the discussion above cites several actual examples that went that far.

21. Indeed, any of the above-described references taken alone or in combination would have provided sufficient suggestion and guidance for a POSITA to have modified Shribman to improve the user experience as well as system efficiency by prioritizing devices for task assignment based on their current workload or idling state, and would have taught them ample techniques in order to do so. It can be seen from the references that this is what many others did when confronting like situations, and their patents and publications shared that solution for those such as the '044/'866 claimed inventors to use.

22. A claim has also been added to using an intermediate proxy device to achieve anonymity as between a requesting client and a tunnel device. First, I do not see more than a limited reflection of this in the disclosure (Fig. 13), and what is disclosed is ineffective because the proxy is identified with the client device. Second, even if there were support for placing such a proxy anywhere, this is already disclosed in well-known anonymity systems such as TOR, whose underlying "onion routing" architecture is reflected in Ex. 1105, Reed. Reed discloses content retrieval through a chain of proxies where anonymity is maintained between the requesting client and each of the intermediate (tunnel) devices, thereby meeting this limitation, for example where there is more than one intermediate node, which is certainly provided for in Reed.

23. In signing this declaration, I recognize that the declaration will be filed as evidence in a contested case before the Patent Trial and Appeal Board of the United States Patent and Trademark and that I may accordingly be subject to cross-examination. If such cross-examination is required of me, I will appear for cross-examination within the United States during the time allotted for cross-examination.

24. This declaration represents only those opinions I have formed to date. I reserve the right to revise, supplement, and/or amend my opinions stated herein based on new information and on my continuing analysis of the materials produced in this proceeding.

I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Executed at McMinnville, Oregon April 21, 2022.

KEITH J. TERUYA